OXFORD

Briefings in Bioinformatics, 2023, 1–10
https://doi.org/10.1093/bib/bbad135
Problem Solving Protocol

UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity

Zhenjiao Du, Xingjian Ding, Yixiang Xu and Yonghui Li

Corresponding author: Yonghui Li, Department of Grain Science and Industry, Kansas State University, Manhattan, KS 66506, USA. Tel.: +1-785-532-4061. E-mail: yonghui@ksu.edu

Abstract

Identification of potent peptides through model prediction can reduce benchwork in wet experiments. However, the conventional process of model buildings can be complex and time consuming due to challenges such as peptide representation, feature selection, model selection and hyperparameter tuning. Recently, advanced pretrained deep learning-based language models (LMs) have been released for protein sequence embedding and applied to structure and function prediction. Based on these developments, we have developed UniDL4BioPep, a universal deep-learning model architecture for transfer learning in bioactive peptide binary classification modeling. It can directly assist users in training a high-performance deep-learning model with a fixed architecture and achieve cutting-edge performance to meet the demands in efficiently novel bioactive peptide discovery. To the best of our best knowledge, this is the first time that a pretrained biological language model is utilized for peptide embeddings and successfully predicts peptide bioactivities through large-scale evaluations of those peptide embeddings. The model was also validated through uniform manifold approximation and projection analysis. By combining the LM with a convolutional neural network, UniDL4BioPep achieved greater performances than the respective state-of-the-art models for 15 out of 20 different bioactivity dataset prediction tasks. The accuracy, Mathews correlation coefficient and area under the curve were 0.7–7, 1.23–26.7 and 0.3–25.6% higher, respectively. A user-friendly web server of UniDL4BioPep for the tested bioactivities is established and freely accessible at https://nepc2pvmzy.us-east-1.awsapprunner.com. The source codes, datasets and templates of UniDL4BioPep for other bioactivity fitting and prediction tasks are available at https://github.com/dzjxzyd/UniDL4BioPep.

Keywords: protein language model, bioactive peptide, protein sequence classification, deep learning, universal architecture

INTRODUCTION

Bioactive peptides (BPs) are protein fragments with positive biological effects, which can be produced from sustainable food protein sources through enzymatic hydrolysis or fermentation [1, 2]. BPs have gained tremendous attention from both researchers and consumers, which is driven by the increasing demand for natural nutraceuticals, concerns of synthetic products, sustainability and, most importantly, the diverse bioactivities exhibited by BPs and their potential to relieve the health burdens [3, 4]. The global BPs market, excluding peptide drugs, is expected to double from 48.6 billion USD in 2020 to 95.7 billion USD by 2028 (https://www.verifiedmarketresearch.com/product/bioactive-peptides-market/). Besides, BP-related research publications were tripled in the past ten years, and thousands of BPs have been identified and deposited in free publicly accessible databases (e.g. DFBP, BIOPEP-UWM, $CAMP_{R3}$, AHTPDB, SpirPep, etc.) [2, 5–11]. So far, some peptides have been commercialized for applications such as drugs, nutraceuticals and cosmeceuticals [2].

Conventionally, biochemistry approaches, including protein pretreatments, protein hydrolysis (or microbial fermentation), hydrolyzate fractionation and purification, and *in vitro* or *in vivo* evaluation are the mainstreams in the novel BPs identification [1, 12–16]. However, these technical routes are hindered by their low efficiency, high cost and heavy reliance on advanced instruments and skilled personnel. To overcome these limitations, some researchers have turned to employ bioinformatics tools, particularly quantitative structure–activity relationships (QSAR) modeling, to fully utilize accumulated BPs data and conduct QSAR models for predicting BPs and decision-making before wet bench research [15, 17–19].

There are three essential steps in QSAR modeling: BPs data collection, peptide representation/encoding/feature extraction

Zhenjiao Du is a Ph.D. student in the Department of Grain Science and Industry at Kansas State University. His research interest is valorization of agricultural by-products by integrating bioinformatic approaches and wet chemistry experiments for bioactive peptide discovery and bioactive protein hydrolysate design. Current studies focus on employing cutting-edge machine-learning strategies and computational chemistry approaches for practical and effective high-throughput bioactive peptide screening protocol development.

Xingjian Ding is a Master student in the Department of Computer Science at Kansas State University. His research focuses on deep learning, bioinformatics, high-performance/parallel computing, and machine learning systems.

Yixiang Xu is a Supervisory Research Food Technologist and Research Leader of the Healthy Processed Foods Research Unit of Agricultural Research Service (ARS) at the United States Department of Agriculture (USDA). Prior to joining USDA-ARS, Dr. Xu has served as a tenured full professor of Food Science at Agriculture Research Station of Virginia State University. Her research interests include development of novel technologies for processing, utilization, and value addition to crops and waste products while addressing sustainability, safety, security and human health challenges.

Yonghui Li is an Associate Professor in the Department of Grain Science and Industry at Kansas State University. His research focuses on the structure, chemistry, modification, and functionality of grain proteins and bioactive peptides with the aim of developing high-quality, functional grain-based foods, ingredients, and nutraceuticals. He and his team employ and integrate wet chemistry, engineering principles, computational simulation, and applied machine learning to deepen the knowledge and advance the research on food proteins and bioactive peptides.

Received: January 19, 2023. **Revised:** February 28, 2023. **Accepted:** March 16, 2023

and model development. The most challenging part is the peptide representation, where peptide sequences are converted into numerical vectors or matrix by different descriptors for further model development [5]. Up to now, various encoding approaches have been proposed, including the physicochemical properties and biochemical properties of amino acids (e.g. AAIndex), amino acid descriptors (e.g. z-scale), amino acid composition, molecular descriptors and fingerprints of the peptide (e.g. three-dimensional (3D) structure information), composition-based descriptors (e.g. amino acid composition and dipeptide composition), binary profile (one-hot encoding), etc. [17, 18, 20–22]. There are some disadvantages in the application of these encoding methods. First, the simple combination of these descriptors leads to high-dimensional features, which include redundant information, hence undermining the predictive performance. Even though many feature selection methods are available to solve the high-dimension problem, the process is time-consuming and tedious because of enormous trial-and-error tests [23–25]. Additionally, the descriptors may not accurately represent peptides. For example, composition-based descriptors lack information about the sequential order, and physicochemical properties can only provide information that can be determined.

In natural language processing (NLP) tasks, word embeddings have been prevalently used to capture semantic properties and linguistic relationships between words and numerical output representations of raw text data for further machine learning (ML) model development [26]. The protein sequence is highly similar to human beings' natural languages, where different amino acids compose a 'language of life' [27]. Most recently, such pretrained deep learning-based language models (LMs) for protein sequence embeddings have been released, such as evolutionary scale modeling (ESM), unified representation (UniRep) and ProtTrans [27–32]. These LMs are trained on large datasets (billions of sequences) to internalize the information encrypted in protein sequences and have exhibited enormous potential in descriptive representations (embeddings) of proteins only relying on their sequences for comparable or improved predictive power in downstream tasks, such as subcellular location, structure prediction and function prediction [27, 30, 31]. Very recently, based on ESM-2 LM, the Meta Fundamental AI Research Protein Team (FAIR) developed a protein 3D structure predictor (ESMFold). Using the similar fold blocks to AlphaFold2, ESMFold was up to 60 times faster than AlphaFold2 and had lower template modeling score (TM-score) in CAMEO and CASP14 datasets [30].

Given that the building blocks of peptides (i.e. amino acids) are the same as those in proteins, and that their bioactivity is comparable to the functionality of proteins, peptide sequences can also be embedded using LMs for bioactivity prediction. To the best of our knowledge, there are few attempts to employ these cutting-edge LMs for peptide embeddings in bioactivity prediction. Transfer learning relying on deep learning has been used to simplify efforts in model architecture design when new tasks can be solved by the previous knowledge [27, 33]. Convolutional neural network (CNN) exhibited promising performance in BPs prediction [17, 18, 25, 34]. Since each element in the peptide embeddings vector/matrix is unique under the specific peptide sequence and the output feature dimension of ESM is a fixed length, it is not necessary to conduct additional feature extraction and processing to unify the feature dimension after embeddings. The fixed feature dimension makes it more suitable for the designing of feature extraction layers in CNN and applying the CNN model to other BP datasets. Therefore, the combination of those LMs and deep learning models may build a universal modeling architecture for BPs prediction among different BP datasets.

The objective of this study was to build a model architecture that relies on a pretrained LM and a CNN model and to investigate the potential of this architecture for transfer learning in different BPs prediction tasks. A state-of-the-art (SOTA) LM, ESM-2, was first used for peptide embeddings, and then a CNN model was proposed to be compatible with the peptide embeddings for BPs prediction (Figure 1). The universal architecture UniDL4BioPep was used to fit 20 different BPs' benchmark datasets, and the performance of the model was compared with the SOTA performance. The UniDL4BioPep has the potential to be used to fit various BP datasets with different bioactivities and generate high-performance models for prediction tasks. It could also inspire future prediction model development for bioactivity prediction.

MATERIALS AND METHODS

Benchmark datasets

All the benchmark datasets were retrieved from the reported SOTA models to conduct a fair and unbiased performance evaluation and comparison. Twenty BPs datasets for eighteen different bioactivities were collected, including angiotensin-converting enzyme (ACE) inhibitory activity (anti-hypertension) [35], dipeptidyl peptidase IV (DPPIV) inhibitory activity (antidiabetes) [21], bitter [25], umami [36], antimicrobial activity [6], antimalarial activity [37], quorum-sensing (QS) activity [38], anticancer activity [39], anti-methicillin-resistant *S. aureus* (MRSA) strains activity [40], tumor T cell antigens (TTCA) [22], blood–brain barrier [41], antiparasitic activity [42], neuropeptide [43, 44], antibacterial activity [45], antifungal activity [45], antiviral activity [45], toxicity [46] and antioxidant activity [18]. The dataset information is summarized in Table 1, and further details on the benchmark datasets can be found in previous studies (also accessible at https://github.com/dzjxzyd/UniDL4BioPep).

Language model for peptide embeddings and data processing

ESM is an LM project initiated by FAIR in 2019 (https://github.com/facebookresearch/esm). Most recently, an updated version of ESM was released as ESM-2, which was trained on the UR50/D 2021_04 dataset and outperformed across a range of structure prediction tasks [30]. ESM-2 contained 6 LMs varying from 48 layers with 15 billion parameters for 5120 output embeddings dimensions to 6 layers with 8 million parameters for 320 output embeddings dimensions. Due to the relatively small size of BPs datasets, the LM (esm2_t6_8M_UR50D) with the lowest output embeddings dimension (320) was selected for peptide embeddings to simplify the CNN model architecture and avoid the curse of dimensionality in model development.

The dataset splitting followed the previous reports where the training and test datasets samples were the same as those used in the SOTA models. In each bioactivity benchmark dataset test, each peptide sequence was first input into the pretrained ESM model to generate a 1*320 dimension numerical vector (Figure 1). Before the CNN model development, a min-max normalization was conducted relying on a training dataset to scale the features in range (0, 1). Normalization in the test dataset is based on the maximum values and minimum from the training dataset. To understand the effectiveness of the ESM-2 in peptide embeddings for bioactivity prediction, uniform manifold approximation and projection (UMAP) was used to visualize the high-dimension embeddings in a two-dimension graph [47]. Besides, visualization based on *t*-distributed stochastic neighbor embedding (t-SNE) was also conducted and provided [48].

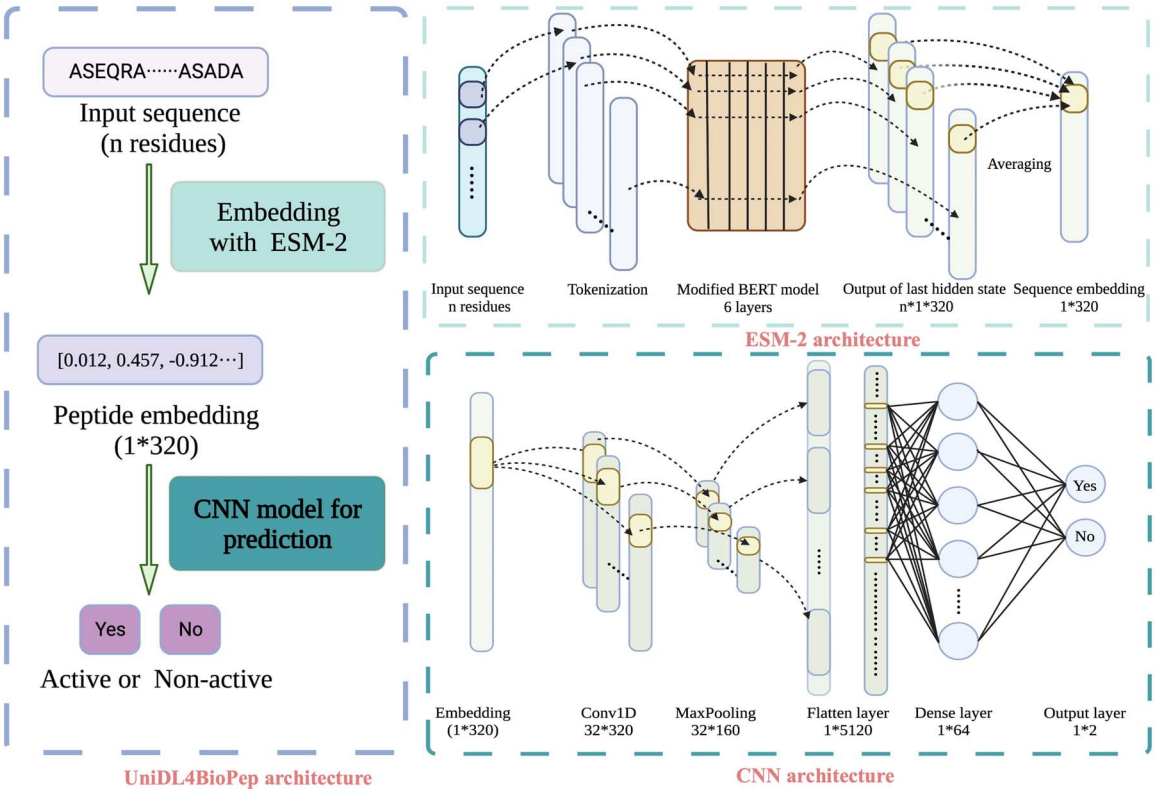

Figure 1. Schematic framework of UniDL4BioPep for peptide bioactivity prediction by integrating ESM-2 and CNN. Note: Peptide of any length is encoded as 320 dimensions embedding by ESM-2, and the embedding is fed into the CNN model. The first layer has 32 filters, and each of these filters undergoes 1D convolution with kernel size 3, stride 1 and ReLU activation function and further down-sampled via a max pooling layer with stride 2. The output 32*160 feature matrix is transformed as a 1D vector in the flatten layer and loaded into a dense layer (fully connected layer) with one hidden layer (64 neurons and ReLU as activation function). The final output layer has two neurons with SoftMax function for prediction.

Table 1. Benchmark datasets collected from publications with state-of-the-art models

Bioactivity	Training dataset	Test dataset	Reference
ACE inhibitory activity	913 Positives and 913 negatives	386 Positives and 386 negatives	[35]
DPP IV inhibitory activity	532 Positives and 532 negatives	133 Positives and 133 negatives	[21]
Bitter	256 Positives and 256 negatives	64 Positives and 64 negatives	[25]
Umami	112 Positives and 241 negatives	28 Positives and 61 negatives	[36]
Antimicrobial activity	3876 Positives and 9552 negatives	2584 Positives and 6369 negatives	[6]
Antimalarial activity	Main dataset (111 positives and 1708 negatives); alternative dataset (111 positives and 542 negatives)	Main dataset (28 positives and 427 negatives); alternative dataset (28 positives and 135 negatives)	[37]
Quorum sensing activity	200 Positives and 200 negatives	20 Positives and 20 negatives	[38]
Anticancer activity	Main dataset (689 positives and 689 negatives); alternative dataset (776 positives and 776 negatives)	Main dataset (172 positives and 172 negatives); alternative dataset (194 positives and 194 negatives)	[39, 56]
Anti-MRSA strains activity	118 Positives and 678 negatives	30 Positives and 169 negatives	[40]
Tumor T cell antigens	470 Positives and 318 negatives	122 Positives and 75 negatives	[22]
Blood–Brain Barrier	100 Positives and 100 negatives	19 Positives and 19 negatives	[41]
Antiparasitic activity	255 Positives and 255 negatives	46 Positives and 46 negatives	[42]
Neuropeptide	1940 Positives and 1940 negatives	485 Positives and 485 negatives	[43, 44]
Antibacterial activity	6583 Positives and 6583 negatives	1695 Positives and 1695 negatives	[45]
Antifungal activity	778 Positives and 778 negatives	215 Positives and 215 negatives	[45]
Antiviral activity	2321 Positives and 2321 negatives	623 Positives and 623 negatives	[45]
Toxicity	1642 Positives and 1642 negatives	290 Positives and 290 negatives	[46]
Antioxidant activity	582 Positives and 541 negatives	146 Positives and 135 negatives	[18]

CNN model architecture

The CNN model was built with Keras framework (http://www.keras.io). For this study, there are totally eight layers (Figure 1) in the CNN models. The first layer was the input layer, where the peptide sequence information was represented by a numerical vector generated by the LM model. Then, the next layer is a 1D convolutional layer with filter sizes of 32, kernel size of 3 and ReLU activation, and then a 1D max pooling layer was added

to reduce the dimensionality. In addition, batch normalization and dropout (rate = 0.15) were added to avoid overfitting. Subsequently, the output of the convolutional layer was flattened and followed by a dense layer containing 64 hidden neurons with ReLU activation function and a dropout rate of 0.15. The last layer is also a dense layer with two neurons with a SoftMax activation function for binary classification. Stochastic gradient descent (SGD) was chosen as the optimizer to accelerate the model fitting and maximize its performance, and step decay was set for the learning rate during the epoch. Besides, an early stop was also set, and the best weight would be stored under the consideration of validation accuracy. The largest epoch time is 200, but in practice, it usually stops around 100 epochs, which takes around two to three minutes on the Google Colab platform.

Model evaluation

Accuracy (ACC), balanced accuracy (BACC), sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC) and area under the curve (AUC) were adopted to evaluate the model performance. Those parameters were calculated based on the number of true positive (TP), false positive (FP), false negative (FN) and true negative (TN). They are calculated by the following equations:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN},$$

$$Sn = \frac{TP}{TP + FN},$$

$$Sp = \frac{TN}{TN + FP},$$

$$BACC = 0.5 * Sn + 0.5 * Sp,$$

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

The AUC is calculated relying on the sklearn 'roc_auc_score' function through the portability distribution of the model prediction in the test dataset.

UniDL4BioPep-FL for imbalanced datasets

In some cases, the bioactive peptide dataset may not be well balanced, resulting in one class being under-represented, which can cause the model in learning less from the minority class [49]. Therefore, the focal loss (FL) function was introduced to tackle the challenge in the imbalanced dataset relying on its ability to down-weight easy examples and thus focus training on hard negatives [50]. Besides, the FL function allows for adjustment of the weighting factor to balance the importance of positive/negative samples. The modified version is named UniDL4BioPep-FL, which contains two hyperparameters, the focusing parameter (gamma) and weighting factor (alpha), which need to be tuned and specified compared to the original UniDL4BioPep model. To be compatible with FL function, the last layer was changed to a one neuron with a sigmoid activation function for binary classification. The equation for FL calculation is below:

$$Focal\ loss = \begin{cases} -\alpha(1-p)^{\gamma}\log(p) & if\ y = 1 \\ -(1-\alpha)\,p^{\gamma}\log(1-p) & otherwise \end{cases}$$

where P is the model's estimated probability for the class with label $y = 1$ (positive); γ is the focal parameter ($\gamma \geq 0$) and α is the weighting factor.

RESULTS AND DISCUSSION

Peptide embeddings analysis

The ESM-2 LMs were trained with the masked language objective, which enabled the LMs to learn dependencies between residues and internalize sequence patterns during millions of repetitions of evolutionarily diverse protein sequences [30]. The architecture of EMS-2 LMs is based on bidirectional encoder representations from transformers (BERT) style encoder, which is known for its ability to return different representations for the same words depending on the contexts [51]. The residues in a peptide sequence contain both information about the residues and information about their position/sequential information. One-hot encoding is a solution to capture both types of information encoded in the peptide sequence. However, the feature matrix is a sparse matrix, and the matrix shape is dependent on the length of the peptide [18]. This can limit the performance of the model since capturing the feature in a spare matrix is difficult. ESM-2 LMs, on the other hand, can generate fixed-length peptide embeddings and dense matrices.

UMAP is a general dimension reduction technique for visualization purposes. Similar to t-SNE, it prioritizes the preservation of local distances over global distances and has the ability to handle outliers and nonlinear relationships, resulting in better performance than principal component analysis in biological data processing [47]. In this study, UMAP distribution of positive and negative samples in both training and testing datasets was plotted in two-dimensional feature space by using 320-dimensional embeddings. As shown in Figure 2, most positive and negative samples are clearly distributed in two clusters. Such distinct differences in the dimensional feature space are generally an indication of the great representation of peptide sequence information for further model development and correspond to optimal models [35, 38]. This is also consistent with UniDL4BioPep's performance in these datasets (Table 2). There is an apparent discrepancy in the number of positive and negative samples in several datasets (Figure 2E, F, K), which is due to an imbalance in the original datasets. Similar observations were also noticed in the t-SNE distribution graphs (Supplementary Figure S1). However, the clustering boundary in t-SNE is more ambiguous compared to UMAP. This may be because of the preservation of more global structures in t-SNE, or it may not be suitable for visualizing clusters in these cases, as has been reported in real-world datasets [47, 52].

Performance evaluation of UniDL4BioPep with SOTA models on the independent test datasets of BPs

To evaluate the robustness and accuracy of our proposed UniDL4BioPep in different BPs datasets, the performance of the SOTA models using the same datasets for training and performance evaluation was collected and compared with the performance of UniDL4BioPep (Table 2). There was a total of twenty datasets for the eighteen bioactivities, where both antimalarial and anticancer activities had two datasets. Among the twenty datasets (Table 2), UniDL4BioPep achieved better performance than the SOTA models in fifteen datasets, where the ACC, MCC and AUC were 0.7–7, 1.23–26.7 and 0.3–25.6% higher than those in the SOTA models. For the remaining datasets, the performance of UniDL4BioPep was also comparable to that achieved by the SOTA models. It is worth noting that during model development in all these datasets, UniDL4BioPep only performed data loading and waited for the trained model for performance

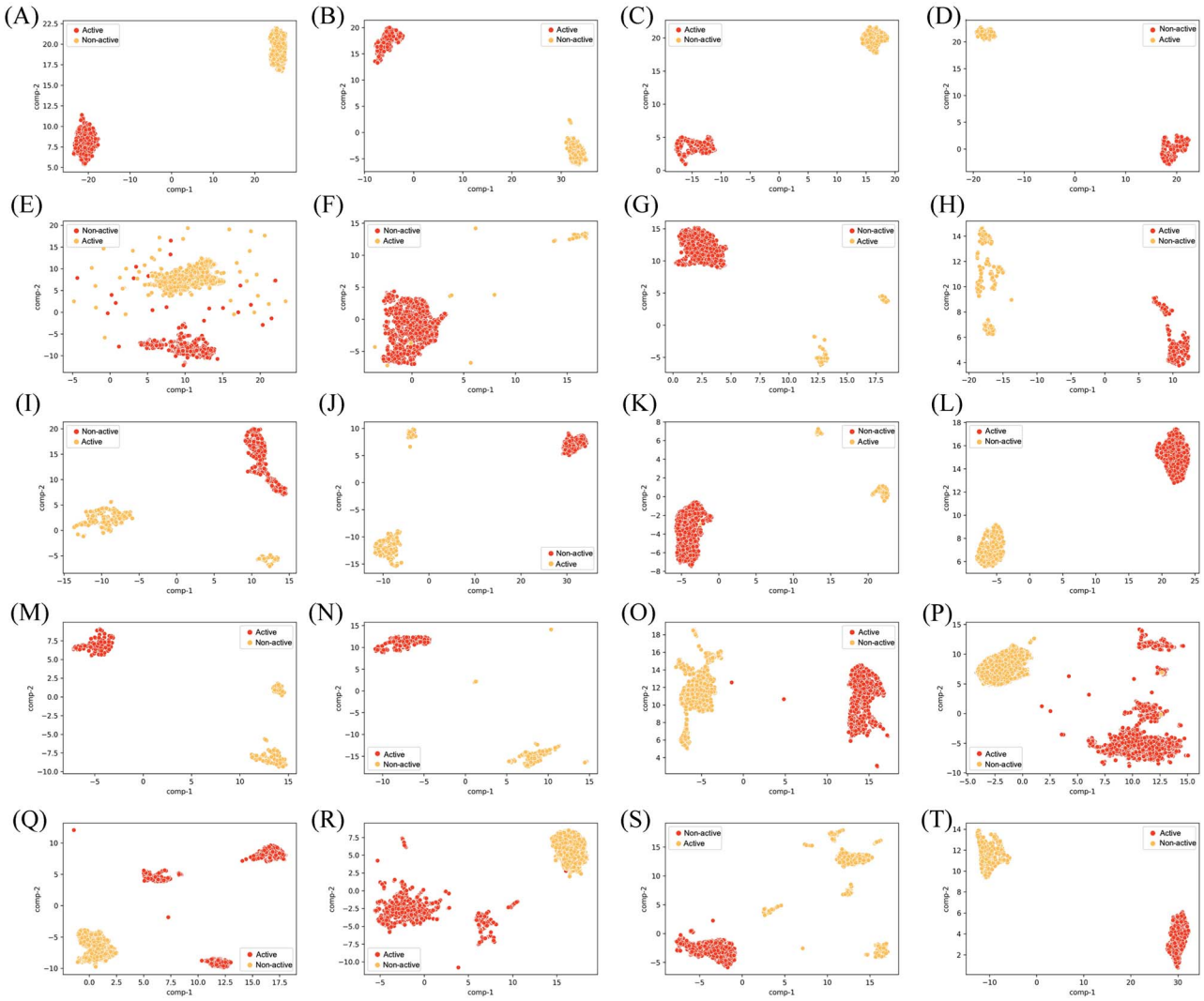

Figure 2. Uniform manifold approximation and projection (UMAP) of positive and negative samples in different bioactive peptide datasets. *Note:* (**A**) ACE inhibitory activity, (**B**) DPP IV inhibitory activity, (**C**) bitter, (**D**) Umami, (**E**) antimicrobial activity, (**F**) antimalarial activity (main dataset), (**G**) antimalarial activity (alternative dataset), (**H**) quorum sensing activity, (**I**) anticancer activity (main dataset), (**J**) anticancer activity (alternative dataset), (**K**) anti-MRSA strains activity, (**L**) tumor T cell antigens, (**M**) blood–brain barrier, (**N**) antiparasitic activity, (**O**) neuropeptide; (**P**) antibacterial activity, (**Q**) antifungal activity, (**R**) antiviral activity, (**S**) toxicity and (**T**) antioxidant activity. Graphs were generated using the whole dataset with n_neighbors parameters = 20.

evaluation in independent test datasets, while each SOTA model had different model architectures, feature selection strategies and hyperparameter tuning. Such performance comparisons demonstrate the great generalization ability of UniDL4BioPep, and its tremendous potential to adapt to new target (i.e. bioactivity or other functions) predictions. The only requirement for the new bioactivity prediction model development is a high-quality and well-labeled dataset.

A total of twenty SOTA model architectures for the respective bioactivities are briefly reviewed and compared with UniDL4BioPep in Table 3. Most SOTA models for the twenty bioactivities adopted similar peptide representation methods, including composition-based methods (e.g. amino acid composition and dipeptide composition), binary profile (one-hot encoding) and physicochemical properties of amino acids (e.g. amino acid index databases). As previously mentioned, these embedding approaches suffer from the loss of sequential information, insufficient description of peptides and the curse of dimensionality. Most researchers still tend to choose traditional ML methods (e.g. random forest) due to the limited data and better explainability compared to neural networks. Therefore, feature selection is essential for simplifying the model complexity and increasing prediction power when the feature dimension is high or too many features are combined. Most recently, Charoenkwan et al. conducted a series of experiments demonstrating better performance and set new SOTA records based on the scoring card method (SCM), which also had a built-in feature selection procedure based on a genetic algorithm [21, 36, 37, 40]. There were also studies employing the idea of word embedding from NLP. In the study of Pang et al., 31 million peptide sequences from the Pfam dataset were used to train a language model by self-learning for peptide embeddings [6]. Other NLP-based approaches, including Word2Vec, TFIDF, Pep2Vec and FastText, were also used in peptide embeddings [25, 43]. UniDL4BioPep achieved comparable performance compared to the SOTA prediction performance in bitterness, neuropeptide and toxicity from previous studies. Specifically, the better performance in bitterness prediction to some extent supports the idea that with

Table 2. Comparison between the performance of UniDL4BioPep and state-of-the-art models from the same benchmark datasets

Bioactivity*	Model name	ACC	BACC	Sn	Sp	MCC	AUC
ACE inhibitory activity	UniDL4BioPep	0.851	0.851	0.836	0.852	0.703	0.901
	mAHTPred [35]	0.883	N/A	0.894	0.873	0.767	0.951
DPP IV inhibitory activity	UniDL4BioPep	0.853	0.853	0.861	0.846	0.707	0.938
	iDPPIV-SCM [21]	0.797	N/A	0.789	0.805	0.594	0.847
Bitter	UniDL4BioPep	0.938	0.938	0.924	0.952	0.875	0.982
	BERT4Bitter [25]	0.922	N/A	0.938	0.906	0.844	0.964
	iBitter-Fuse [54]	0.93	N/A	0.938	0.922	0.859	0.933
Umami	UniDL4BioPep	0.888	0.875	0.846	0.905	0.735	0.948
	UniDL4BioPep-FL (gamma = 3)**	0.888	0.883	0.892	0.875	0.733	0.943
	iUmami-SCM [36]	0.865	N/A	0.714	0.934	0.679	0.898
Antimicrobial activity	UniDL4BioPep	0.962	0.958	0.968	0.948	0.908	0.991
	UniDL4BioPep-FL (gamma = 3.5)**	0.96	0.961	0.96	0.963	0.903	0.991
	TransImbAMP [6]	N/A	0.969	0.963	0.974	N/A	N/A
Antimalarial activity (main dataset)	UniDL4BioPep	0.98	0.989	1	0.979	0.815	0.921
	UniDL4BioPep-FL (gamma = 4)**	0.978	0.965	0.979	0.95	0.793	0.898
	iAMAP-SCM [37]	0.978	0.826	0.654	0.998	0.776	0.82
Antimalarial activity (alternative dataset)	UniDL4BioPep	0.975	0.97	0.962	0.978	0.912	0.987
	UniDL4BioPep-FL (gamma = 1)**	0.989	0.993	0.985	1.0	0.9570	0.987
	iAMAP-SCM [37]	0.957	0.896	0.808	0.985	0.834	0.903
Quorum sensing activity	UniDL4BioPep	0.95	0.955	0.909	1	0.905	0.99
	iQSP [55]	0.93	N/A	N/A	N/A	0.86	0.96
	QSPred-FL [38]	0.943	N/A	0.935	0.95	0.885	0.945
Anticancer activity (main dataset)	UniDL4BioPep	0.735	0.735	0.734	0.737	0.471	0.805
	iACP-FSCM [56]	0.825	0.825	0.726	0.903	0.646	0.812
	AntiCP 2.0 [39]	0.754	0.754	0.774	0.734	0.51	N/A
Anticancer activity (alternative dataset)	UniDL4BioPep	0.946	0.948	0.978	0.918	0.894	0.971
	iACP-FSCM [56]	0.889	N/A	0.876	0.902	0.779	0.93
	AntiCP 2.0 [39]	0.92	N/A	0.923	0.918	0.84	N/A
Anti-MRSA strains activity	UniDL4BioPep	0.99	0.994	1	0.988	0.96	0.999
	UniDL4BioPep-FL (gamma = 3)**	0.994	0.997	0.994	1	0.98	0.999
	SCMRSA [40]	0.96	0.935	0.9	0.97	0.848	0.986
Tumor T cell antigens	UniDL4BioPep	0.751	0.743	0.763	0.724	0.457	0.788
	UniDL4BioPep-FL (gamma = 2)**	0.746	0.762	0.734	0.791	0.446	0.796
	iTTCA-Hybrid [22]	0.71	N/A	0.844	0.493	0.363	0.756
Blood–Brain Barrier	UniDL4BioPep	0.842	0.846	0.882	0.809	0.688	0.992
	BBPpred [41]	0.7895	N/A	0.6316	0.9474	0.6102	0.7895
Antiparasitic activity	UniDL4BioPep	0.891	0.911	0.821	1	0.8	0.94
	PredAPP [42]	0.88	N/A	0.978	0.783	0.775	0.922
Neuropeptide	UniDL4BioPep	0.892	0.892	0.875	0.909	0.784	0.953
	PredNeuroP [44]	0.897	N/A	0.886	0.907	0.794	0.954
	NeuroPred-CLQ [43]	0.936	N/A	0.897	0.975	0.875	0.988
Antibacterial activity	UniDL4BioPep	0.94	0.941	0.966	0.915	0.881	0.978
	ABPDiscover [45]	0.935	N/A	0.912	0.957	0.87	0.975
Antifungal activity	UniDL4BioPep	0.948	0.952	1	0.904	0.902	0.994
	ABPDiscover [45]	0.942	N/A	0.921	0.963	0.884	0.988
Antiviral activity	UniDL4BioPep	0.842	0.853	0.916	0.79	0.694	0.907
	ABPDiscover [45]	0.828	N/A	0.764	0.892	0.662	0.896
Toxicity	UniDL4BioPep	0.941	0.941	0.923	0.959	0.883	0.978
	ATSE [46]	0.952	N/A	0.965	0.94	0.903	0.976
Antioxidant activity	UniDL4BioPep	0.804	0.804	0.81	0.799	0.608	0.872
	AnOxPePred-FRS [18]	N/A	N/A	N/A	N/A	0.48	0.79

Note: *The corresponding datasets are summarized in Table 1; **Only gamma is specified for UniDL4BioPep-FL; performance parameters are marked in bold when UniDL4BioPep achieved better performance; N/A: not available in the original paper. Abbreviations: ACC: accuracy; AUC: area under the curve; BACC: balanced accuracy; Sn: sensitivity; Sp: specificity; MCC: Matthews correlation coefficient.

appropriate feature dimensions, deep learning can also achieve great performance in a small dataset (e.g. a total of 320 positive and 320 negative samples) [53, 54].

In UniDL4BioPep, peptide representation is generated by LM (ESM-2), which takes into account each residue and sequential information to generate fixed-length peptide embeddings. There are also other ESM-2-based LMs with longer fixed-length output dimensions [30]. However, lower dimensions correspond to fewer parameters in the following deep learning model, and thus we chose the available pretrained ESM-2

Table 3. Comparison and overview of UniDL4BioPep and the state-of-the-art models for binary classification in peptide bioactivity

Bioactivity*	Model name	Peptide representation**	Feature selection methods	Model development***	References
Various bioactivities, universal	UniDL4BioPep	Language model, evolutionary scale modeling (ESM-2, esm2_t6_8M_UR50D with 320 fixed length dimension output for any length peptide input)	None	CNN model with six layers	This work
ACE inhibitory activity	mAHTPred	AAC, AAI, BPF, CTD, DPC, OVP, Hybrid features, BPF.	Feature importance scores and sequential forward search (SFS) based on RF	An ensemble model with ERT, RF, SVM, GB	[35]
DPP IV inhibitory activity	iDPPIV-SCM	AAC, DPC	None	SCM	[21]
Bitter	iBitter-Fuse	AAC, DPC, PAAC, APAAC, AAI	Genetic algorithm utilizing self-assessment-report	SVM	[54]
Bitter	BERT4Bitter	TFIDF, Pep2Vec, FastText	None	CNN, LSTM neural network, BERT-based BiLSTM	[25]
Umami	iUmami-SCM	DPC	None	DT, kNN, MLP, NB, SVM and RF, SCM	[36]
Antimicrobial activity	TransImbAMP	Embedded by a self-designed language model	None	MLP	[6]
Antimalarial activity (main dataset and alternative dataset)	iAMAP-SCM	AAC, DPC, TPC, CTD, CTDC, AAI, CTD, CTDD, CTDC, DPS	None	kNN, MLP, SVM and RF, SCM	[37]
Quorum sensing activity	QSPred-FL	AAC, CTD, 188-bit features, GDC, ASDC, IT, BPF, N- and C-terminus approach, 21-bit features, OVP	Minimum redundancy feature selection and SFS based on RF	RF, NB, SVM, LR and J48	[38]
Quorum sensing activity	iQSP	AAI	Genetic algorithm utilizing self-assessment-report	SVM	[55]
Anticancer activity (main dataset and alternative dataset)	iACP-FSCM	AAC, DPC, composition on terminal region	None	Flexible SCM	[56]
Anticancer activity (main dataset and alternative dataset)	AntiCP 2.0	AAC, DPC, BPF, split composition	None	ANN, SVM, KNN, ETree, RF, Ridge classifier	[39]
Anti-MRSA strains activity	SCMRSA	AAC, AAI, APS, DPC, CTD	None	DT, KNN, LR, NB, PLS, SVM, SCM	[40]
Tumor T cell antigens	iTTCA-Hybrid	AAC, DPC, PACC, CTDD, AAI, hybrid feature	None	SVM, RF	[22]
Anti-parasitic	PredAPP	AAC, AAI, CT5, CTD, DPC, GAAC, GDPC, NT5 and PAAC	Nine feature was used to build model for prediction with six modeling methods (KNN, LR, MLP, RF, SVM and XGB). The output of the best model for each feature was used as final selected feature.	KNN, LR, MLP, RF, SVM and XGB	[42]
Blood–brain barrier	BPPred	AAC, CTD, BIT12, BIT118, GDPC, GTPC, GGAP, BIT21, IT, OLP, DPC, ASDC, GGAP, PAAC, CTF, BIT20, QSO	Combination of F1 score obtained from ERT, Spearman correlation coefficient and SFS-based selection relying on ERT, XGB, KNN, LR, MLP, RF and SVM	ERT, XGB, KNN, LR, MLP, RF and SVM	[41]
Neuropeptide	PredNeuroP	AAC, DPC, BPF, AAE, AAAI, GAAC, GDPC, GTPC, CTD	Five classifiers (ANN, ERT, XGB, KNN and LR) were used to build 45 base-learners based on the nine features. Eight of them were selected relying on Pearson correlation coefficients and accuracy.	A stacking model (eight optimal base-learner as the first layer and LR as the second layer)	[44]
Neuropeptide	NeuroPred-CLQ	Word2Vec for sequence embeddings	None	CNN, TCNN and a multi-head attention layer were connected sequentially.	[43]
Antibacterial activity/antifungal activity/antiviral activity	ABPDiscover****	Generated by ProtDCal	Six feature selection methods (correlation subset, Relief-F, information gain, gain ratio and symmetrical uncertainty) were used separately and jointly to generated subset features. Wrapper feature selection using genetic algorithm with RF as learning method.	RF	[45]
Toxicity	ATSE	Molecular graphs and position-specific scoring matrix (PSSM) of sequences	None	GNN and CNN_BiLSTM were used to extract information from molecular graphs and PSSM, respectively. Their output was flattened and loaded into a attention module and output module (fully connected layer).	[46]
Antioxidant activity	AnOxPePred	BF	None	CNN	[18]

Note: *The corresponding datasets are summarized in Table 1, ** and ***feature names and model methods in bold indicate that they were combined to achieve the state-of-the-art performance in Table 3, while in studies with feature selection approach relying on the entire feature pool, all the features are not marked in bold. ****For each bioactivity, feature selection needs to be redone. Abbreviation in peptide representation: AAC: amino acid composition; AAE: amino acid entropy; AAI: amino acid index database; BPF: binary profiles (one hot encoding); APAAC: amphiphilic pseudo-amino acid composition; ASDC: integration of the DPC and GDC; BIT12: combination of SAAC, GAAC, molecular weight and peptide length; BIT 118: combination of AAC and CTD; BIT20: same as BPF; BIT 21: same as TOB; CTF: conjoint triad feature; CTD: composition-transition-distribution; CTDC: composition-transition-distribution composition; CTDD: composition-transition-distribution distribution; CTDT: composition-transition-distribution transition; CKSAAGP: composition of k-spaced amino acid group pairs; DPC: dipeptide composition; APS: amino acid propensities; DPS: propensities of 400 dipeptides; Estate: electrotopological state atom types; FP4: presence of SMARTS patterns for functional groups; GAAC: grouped amino acid composition; GDC: the correlation of nonadjacent residue pairs; GDPC: grouped dipeptide composition; GTPC: grouped tripeptide composition; GGAP: g-gap dipeptide composition; IT: the input peptide is encoded as 3-dimensional vector (Shannon entropy, relative Shannon entropy and information gain score); MACCS: binary representation of chemical features defined by MACCS key; OVP: overlapping property features; OF: other features (absolute charge per residue, aliphatic index, a fraction of transforming residue, molecular weight and sequence length); OLP: overlapping property feature; PAAC: pseudo amino acid composition; Pubchem: binary representation of substructures defined by PubChem; QSO: quasi-sequence order feature; SAAC: selected amino acid composition; TFIDF: term frequency-inverse document frequency; TOB: twenty-one-bit features; TPC: tripeptide composition Abbreviation in model development: ANN: artificial neural network; BERT: bidirectional encoder representations from transformers; BiLSTM: bidirectional long-short term memory; CNN: convolutional neural network; DT: decision tree; ERT: extremely randomized tree; ETree: extra trees; GB: gradient boosting; GNN: graph neural networks; kNN: k nearest neighbors; LR: logistic regression; LSTM: long-short term memory; MLP: multilayer perceptron; NB, naïve Bayes, PLS: partial least squares regression; RF: random forest; SCM: scoring card method; SVM: support vector machine; TCNN: temporal convolutional neural network; XGB: eXtreme gradient boosting

with the lowest output dimension (320 dimensions). ESM-2 uses bidirectional encoder representations from transformers (BERT)-based architecture with modifications. The output is the hidden state of the ESM-2 model, which is tuned by millions of protein sequences. Hence, both the value of each element in the output vector and the relationship between each element are important in representing the peptide sequence and the loss of each element might undermine the information contained in the peptide embeddings. As a result, feature selection was not conducted. The CNN architecture is inspired by the CNN architectures in literature and also practical application of CNN in the handwritten digits recognition MNIST database with a few modifications based on the dimension of peptide embeddings from ESM-2 [18, 25]. The shallow CNN model in UniDL4BioPep had fewer parameters and was compatible with embedding dimensions from the ESM-2. BPs with small datasets (e.g. bitter, umami, quorum sensing, tumor T cell antigens, antiparasitic activity, etc.) had improved performance in test dataset prediction with UniDL4BioPep, compared to the SOTA models. Besides, a modified universal model version, UniDL4BioPep-FL, is introduced to meet the needs of imbalanced dataset modeling relying on the FL function. Improvement in the balance of Sn and Sp was observed in six imbalanced benchmark datasets (umami, antimicrobial activity, antimalarial activity (main and alternative datasets), anti-MRSA strains activity and tumor T cell antigens), which supports its validity. Overall, UniDL4BioPep is a universal architecture in BPs prediction and overcomes the challenges in dataset size (compatible), peptide representation (repeatable and fixed approach), feature selection (not needed), ML methods selection (not needed) and hyperparameter tuning (universal to the fixed-length feature dimension, only needed for two hyperparameters in UniDL4BioPep-FL).

CONCLUSION

Novel bioactive peptide exploration boosted by bioinformatics can significantly reduce experimental periods and cost. As such, fast and accurate prediction models are highly desirable. In this study, we first introduced the latest pretrained language model based on millions of protein sequences for peptide embeddings and further application in bioactivity prediction. Besides, relying on the fixed-length output of the embeddings, corresponding convolutional layers and dense layers were designed for feature extraction of the embeddings and bioactivity prediction. Because of the self-learned characteristics of CNN models, it is possible to build a universal architecture for general bioactivity prediction, and thus UniDL4BioPep was proposed. The model was trained and evaluated on eighteen different bioactivities with a total of twenty datasets and was compared with the state-of-the-art models. The UMAP results showed that the pretrained language model (esm2_t6_8M_UR50D) is suitable for peptide embeddings. The CNN model architecture was compatible with the embeddings and exhibited better or comparable performance than the state-of-the-art models, which further confirmed the feasibility and superiority of using a language model in peptide embeddings over conventional approaches.

UniDL4BioPep has great potential to be applied to other bioactive peptide predictions and generate prediction models with comparable performance. Users would only need to prepare their BP dataset in MS Excel in a required format (the first column for sequence and the second column for label). Then, UniDL4BioPep can automatically read the dataset, embed the peptide sequences, split datasets (8:2 ratio as train and test datasets), generate a prediction model, evaluate the model performance and save the model for further prediction needs. Besides, we also provide a modified version (UniDL4BioPep-FL) for advanced usage of the architecture for imbalanced datasets, where two hyperparameters are needed to be tuned. The templates for UniDL4BioPep-based model training for other BP datasets and the employment of generated model for prediction are available at https://github.com/dzjxzyd/UniDL4BioPep. Besides, UniDL4BioPep/UniDL4BioPep-FL can also be used for multiclass classification model development, which further expands its application scenario. All the scripts were written and tested on Google Colab, which allows users to execute them through browsers.

With the great predictive performance of UniDL4BioPep, a user-friendly web server for accelerating peptide screening and practical applications has been deployed and is freely available online at https://nepc2pvmzy.us-east-1.awsapprunner.com/. Users can choose a prediction model and submit a peptide sequence, a batch of sequences or a file in FASTA, txt, Microsoft Excel formats to the webserver and will receive the predicted results in real time. We anticipate that the proposed UniDL4BioPep architectures will be an efficient and powerful tool for BP discovery and inspire future model design.

Key Points

- The UniDL4BioPep architecture outperformed the state-of-the-art models in 15 out of 20 peptide bioactivity datasets.
- A user-friendly web server was developed for eighteen bioactivity predictions of any length peptides with outstanding prediction performance from the UniDL4BioPep architecture.
- UniDL4BioPep has the potential to be applied to fit various bioactive peptide datasets and is expected to achieve cutting-edge performance.
- The advanced protein language model, ESM-2, exhibited great power to embed peptide sequences with any length.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

DATA AND CODE AVAILABILITY

The datasets and source codes used in this study are available at https://github.com/dzjxzyd/UniDL4BioPep.

FUNDING

This is contribution No. 23-192-J from the Kansas Agricultural Experimental Station. This work was supported in part by the Agriculture and Food Research Initiative Competitive Grant no. 2020-68008-31408 and no. 2021-67021-34495 from the USDA National Institute of Food and Agriculture and a seed grant from the Global Food Systems initiative of Kansas State University.

REFERENCES

1. Ulug SK, Jahandideh F, Wu J. Novel technologies for the production of bioactive peptides. *Trends Food Sci Technol* 2021;**108**:27–39.

2. Du Z, Li Y. Review and perspective on bioactive peptides: a roadmap for research, development, and future opportunities. *J Agric Food Res* 2022;**9**:100353.
3. FitzGerald RJ, Cermeño M, Khalesi M, *et al.* Application of in silico approaches for the generation of milk protein-derived bioactive peptides. *J Funct Foods* 2020;**64**:103636.
4. Iwaniak A, Darewicz M, Mogut D, Minkiewicz P. Elucidation of the role of in silico methodologies in approaches to studying bioactive peptides derived from foods. *J Funct Foods* 2019;**61**:103486.
5. Du Z, Li Y, Comer J. Bioinformatics approaches to discovering food-derived bioactive peptides: reviews and perspectives. *Trends Anal Chem* 2023. Under Review.
6. Pang Y, Yao L, Xu J, *et al.* Integrating transformer and imbalanced multi-label learning to identify antimicrobial peptides and their functional activities. *Bioinformatics* 2022;**38**:5368–74.
7. Minkiewicz I, Darewicz. BIOPEP-UWM database of bioactive peptides: current opportunities. *IJMS* 2019;**20**:5978.
8. Waghu FH, Barai RS, Gurung P, Idicula-Thomas S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res* 2016;**44**:D1094–7.
9. Kumar R, Chaudhary K, Sharma M, *et al.* AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides. *Nucleic Acids Res* 2015;**43**:D956–62.
10. Anekthanakul K, Hongsthong A, Senachak J, Ruengjitchatchawalya M. SpirPep: an in silico digestion-based platform to assist bioactive peptides discovery from a genome-wide database. *BMC Bioinf* 2018;**19**:149.
11. Qin D, Bo W, Zheng X, *et al.* DFBP: a comprehensive database of food-derived bioactive peptides for peptidomics research. *Bioinformatics* 2022;**38**:3275–80.
12. Wen C, Zhang J, Zhang H, *et al.* Plant protein-derived antioxidant peptides: isolation, identification, mechanism of action and application in food systems: a review. *Trends Food Sci Technol* 2020;**105**:308–22.
13. Barati M, Javanmardi F, Mousavi Jazayeri SMH, *et al.* Techniques, perspectives, and challenges of bioactive peptide generation: a comprehensive systematic review. *Comp Rev Food Sci Food Safe* 2020;**19**:1488–520.
14. Perez Espitia PJ, de Fátima Ferreira Soares N, dos Reis Coimbra JS, *et al.* Bioactive peptides: synthesis, properties, and applications in the packaging and preservation of food. *Comp Rev Food Sci Food Safe* 2012;**11**:187–204.
15. Tu M, Cheng S, Lu W, du M. Advancement and prospects of bioinformatics analysis for studying bioactive peptides from food-derived protein: sequence, structure, and functions. *Trends Anal Chem* 2018;**105**:7–17.
16. Duffuler P, Bhullar KS, de Campos Zani SC, Wu J. Bioactive peptides: from basic research to clinical trials and commercialization. *J Agric Food Chem* 2022;**70**:3585–95.
17. Chen J, Cheong HH, Siu SWI. xDeep-AcPEP: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. *J Chem Inf Model* 2021;**61**:3789–803.
18. Olsen TH, Yesiltas B, Marin FI, *et al.* AnOxPePred: using deep learning for the prediction of antioxidative properties of peptides. *Sci Rep* 2020;**10**:21471.
19. Du Z, Li Y. Computer-aided approaches for screening Antioxidative dipeptides and application to sorghum proteins. *ACS Food Sci Technol* 2022;**2**:1781–8.
20. Kalyan G, Junghare V, Khan MF, *et al.* Anti-hypertensive peptide predictor: a machine learning-empowered web server for prediction of food-derived peptides with potential angiotensin-converting enzyme-I inhibitory activity. *J Agric Food Chem* 2021;**69**:14995–5004.
21. Charoenkwan P, Kanthawong S, Nantasenamat C, *et al.* 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* 2020;**19**:4125–36.
22. Charoenkwan P, Nantasenamat C, Hasan MM, Shoombuatong W. iTTCA-hybrid: improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. *Anal Biochem* 2020;**599**:113747.
23. Du Z, Tian W, Tilley M, *et al.* Quantitative assessment of wheat quality using near-infrared spectroscopy: a comprehensive review. *Comp Rev Food Sci Food Safe* 2022;**21**:2956–3009.
24. Du Z, Wang D, Li Y. Comprehensive evaluation and comparison of machine learning methods in QSAR Modeling of antioxidant tripeptides. *ACS Omega* 2023;**7**:25760–71.
25. Charoenkwan P, Nantasenamat C, Hasan MM, *et al.* BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics* 2021;**37**:2556–62.
26. Wang Y, Liu S, Afzal N, *et al.* A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform* 2018;**87**:12–20.
27. Elnaggar A, Heinzinger M, Dallago C, *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* 2021;**1–1**:1.
28. Alley EC, Khimulya G, Biswas S, *et al.* Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**:1315–22.
29. Rao R, Meier J, Sercu T, *et al.* Transformer protein language models are unsupervised structure learners. Vienna, Austria, 2021. Paper3581. International Conference on Learning Representations, Appleton, WI, USA.
30. Lin Z, Akin H, Rao R, *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. *Science* 2023;**379**:1123–30.
31. Dallago C, Schütze K, Heinzinger M, *et al.* Learned Embeddings from deep learning to visualize and predict protein sets. *Curr Protocol* 2021;**1**:e113.
32. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**:e2016239118.
33. Tammina S. Transfer learning using VGG-16 with deep convolutional neural network for classifying images. *IJSRP* 2019;**9**:9420.
34. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 2018;**34**:2740–7.
35. Manavalan B, Basith S, Shin TH, *et al.* mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 2019;**35**:2757–65.
36. Charoenkwan P, Yana J, Nantasenamat C, *et al.* iUmami-SCM: a novel sequence-based predictor for prediction and analysis of umami peptides using a scoring card method with propensity scores of dipeptides. *J Chem Inf Model* 2020;**60**:6666–78.
37. Charoenkwan P, Schaduangrat N, Lio P, *et al.* iAMAP-SCM: a novel computational tool for large-scale identification of antimalarial peptides using estimated propensity scores of dipeptides. *ACS Omega* 2022;**7**:41082–95.
38. Wei L, Hu J, Li F, *et al.* Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Brief Bioinform* 2020;**21**:106–19.

39. Agrawal P, Bhagat D, Mahalwal M, *et al.* AntiCP 2.0: an updated model for predicting anticancer peptides. *Brief Bioinform* 2021;**22**:bbaa153.
40. Charoenkwan P, Kanthawong S, Schaduangrat N, *et al.* SCMRSA: a new approach for identifying and Analyzing anti-MRSA peptides using estimated propensity scores of dipeptides. *ACS Omega* 2022;**7**:32653–64.
41. Dai R, Zhang W, Tang W, *et al.* BBPpred: sequence-based prediction of blood-brain barrier peptides with feature representation learning and logistic regression. *J Chem Inf Model* 2021;**61**:525–34.
42. Zhang W, Xia E, Dai R, *et al.* PredAPP: predicting anti-parasitic peptides with Undersampling and ensemble approaches. *Interdiscip Sci Comput Life Sci* 2022;**14**:258–68.
43. Chen S, Li Q, Zhao J, *et al.* NeuroPred-CLQ: incorporating deep temporal convolutional networks and multi-head attention mechanism to predict neuropeptides. *Brief Bioinform* 2022;**23**:bbac319.
44. Bin Y, Zhang W, Tang W, *et al.* Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features. *J Proteome Res* 2020;**19**:3732–40.
45. Pinacho-Castellanos SA, García-Jacas CR, Gilson MK, *et al.* Alignment-free antimicrobial peptide predictors: improving performance by a thorough analysis of the largest available data set. *J Chem Inf Model* 2021;**61**:3141–57.
46. Wei L, Ye X, Xue Y, *et al.* ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform* 2021;**22**:bbab041.
47. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *J Open Source Softw* 2018;**3**:861.
48. van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008;**9**:2579–605.
49. Lemaıtre G, Nogueira F. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 2017;**18**:1–5.
50. Lin T-Y, Goyal P, Girshick R, *et al.* Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**:318–27.
51. Devlin J, Chang M-W, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Cedarville, OH, USA, 2019*, Vol 1, p. 4171–86. Minneapolis, Minnesota, 2019.
52. Yang Z, Chen Y, Corander J. T-SNE is not optimized to reveal clusters in data. *bioRxiv* 2021; 2110.02573.
53. Charoenkwan P, Yana J, Schaduangrat N, *et al.* iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics* 2020;**112**:2813–22.
54. Charoenkwan P, Nantasenamat C, Hasan MM, *et al.* iBitter-fuse: a novel sequence-based bitter peptide predictor by fusing multi-view features. *IJMS* 2021;**22**:8958.
55. Charoenkwan P, Schaduangrat N, Nantasenamat C, *et al.* iQSP: a sequence-based tool for the prediction and analysis of quorum sensing peptides using informative physicochemical properties. *Int J Mol Sci* 2020;**21**:75.
56. Charoenkwan P, Chiangjong W, Lee VS, *et al.* Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. *Sci Rep* 2021;**11**:3017.