Briefings in Bioinformatics, 00(00), 2021, 1–14

doi: 10.1093/bib/bbab008
Problem Solving Protocol

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

Anticancer peptides prediction with deep representation learning features

Zhibin Lv†, Feifei Cui†, Quan Zou, Lichao Zhang and Lei Xu

Corresponding author: Quan Zou, E-mail: zouquan@nclab.net; Lei Xu, E-mail: csleixu@szpt.edu.cn; Lichao Zhang, E-mail: lczhang5354@szu.edu.cn
†These authors contributed equally to this work.

Abstract

Anticancer peptides constitute one of the most promising therapeutic agents for combating common human cancers. Using wet experiments to verify whether a peptide displays anticancer characteristics is time-consuming and costly. Hence, in this study, we proposed a computational method named identify anticancer peptides via deep representation learning features (iACP-DRLF) using light gradient boosting machine algorithm and deep representation learning features. Two kinds of sequence embedding technologies were used, namely soft symmetric alignment embedding and unified representation (UniRep) embedding, both of which involved deep neural network models based on long short-term memory networks and their derived networks. The results showed that the use of deep representation learning features greatly improved the capability of the models to discriminate anticancer peptides from other peptides. Also, UMAP (uniform manifold approximation and projection for dimension reduction) and SHAP (shapley additive explanations) analysis proved that UniRep have an advantage over other features for anticancer peptide identification. The python script and pretrained models could be downloaded from https://github.com/zhibinlv/iACP-DRLF or from http://public.aibiochem.net/iACP-DRLF/.

Key words: anticancer; peptide; representation learning; light gradient boosting; feature selection

Introduction

Cancer is devastating, as it kills millions of people around the world every year [1–3]. How to treat cancer is a major medical challenge facing mankind. At present, the main methods to treat cancer are radiotherapy, chemotherapy and targeted therapy [4–7]. The idea behind these treatments is to kill cancer cells, but they also damage normal cells [8]. These methods are with obvious side effects and are unable to be afforded by many patients [9]. Anticancer peptides (ACPs) constitute a class of peptides that have been found to have anticancer effects, and are usually characterized by a sequence length of no more than 50 amino acid residues [10]. Using ACPs to treat cancer is a valuable potential alternative to current cancer therapies [11]. The ACPs show some significant advantages over other treatments for cancers: they are safer since they are natural biological inhibitors; and they display higher selectivity toward killing cancer cells due to their natural cationic properties to selectively interact with the anionic cell membrane components of the cancer cell [12]. In recent years, the ACPs therapy has been extensively explored and applied in preclinical settings and different stages of clinical trials against various types of tumors

Zhibin Lv is a senior engineer and a postdoctoral at University of Electronic Science and Technology of China. He received his PhD degree from Peking University in 2013. His research interest is on machine learning for bioinformatics.
Feifei Cui is currently a postdoctoral researcher at the University of Electronic Science and Technology of China. She received her PhD degree from the University of Tokyo, Japan. Her research interests include bioinformatics, deep learning and biological data mining.
Quan Zou is a professor in Institute of Fundamental and Frontier Sciences at University of Electronic Science and Technology of China. He received his PhD degree from Harbin Institute of Technology in 2009. He focuses on area of bio-sequences analysis.
Lichao Zhang is a lecturer at the School of Intelligent Manufacturing and Equipment, Shenzhen Institute of Information Technology. Her research interests include machine learning and bioinformatics.
Lei Xu is an associate professor at the School of Electronic and Communication Engineering, Shenzhen Polytechnic. Her research interests include machine learning and bioinformatics.
Submitted: 6 October 2020; **Received (in revised form):** 20 December 2020

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

[13, 14]. Although the ACPs are safer than traditional broad-spectrum drugs and have become a more competitive treatment option than small molecules and antibodies, very few ACPs have actually been found. According to data collected by CancerPPD, a professionally maintained database of the ACPs, no >3000 ACP sequences have been experimentally verified, and if the ACPs with high sequence similarity are removed, only a few of them are clinically available [15].

Given the therapeutic advantages of the ACPs and the relatively few experimentally validated ACPs, it is critical to develop computational methods for identifying the ACPs from the non-anticancer peptides (non-ACPs) [16, 17], especially for large-scale protein/peptides sequences generated by next-generation sequencing technology. In the past few years there has been an emergence of computational predictions of the ACPs based on machine learning [18]. These methods typically transform the amino acid sequences of anticancer and non-ACPs into various numerical features, and then machine learning algorithms are used to learn patterns from these features to distinguish the ACPs from the non-ACPs [11, 16, 18–29]. Common feature extraction methods include the use of dipeptide composition (DPC) and binary profiles implemented in the web server AntiCP developed by Tyagi *et al.* [15], Chou's pseudo-amino acid composition (PseAAC) used by Hajisharifi *et al.* [19], protein relatedness measurement parameters in the web server ACPP developed by Vijayakumar and Lakshmi [20], PseAAC with the G-Gap dipeptide mode in the sequence-based tool iACP developed by Chen *et al.* [21], amino acid composition (AAC) and physicochemical properties in MLACP developed by Manavalan *et al.* [23], k-gram in the web server AMPFun developed by Chung *et al.* [30] and others [22]. In recent years, Wei *et al.* used feature selection and feature fusion methods to develop ACPred-FL [24], PEPred-suite [26] and ACPred-Fuse [25] with better accuracy. In 2020, a new and better method named AntiCP 2.0 emerged [16]. In contrast to the classical computational learning methods mentioned above, such as support vector machine and random forest (RF), Yi *et al.* proposed the use of a long short-term memory (LSTM) neural network model in the ACP prediction method ACP-DL [27]. Sequence feature extraction plays an important role in the prediction of biological sequences [31]. Limited by the short sequences of ACPs, it is not enough to just extract more sequence statistical information for current predictors [25, 27]. Therefore, although these predictors have greatly promoted the research of predicting ACPs, it is still necessary to develop higher-performance ACPs prediction methods.

In the last few years, inspired by natural language processing using deep representation learning [32, 33] and transfer learning [34], sequence-based deep representation learning for proteins and peptides have been emerging [35–42]. Typical examples of such methods include ProFET, ProVec, unified representation (UniRep), TAPE, ProGen and UDSMProt, which have been shown to be powerful tools for protein function prediction, protein structure prediction, reasonable protein design, protein–protein interaction and GO prediction [36, 43–56]. The methods are usually based on an unsupervised or a semi-supervised training learning by using extremely large data sets such as UniRef50 [57] and the Pfam protein families database [58], which include tens of millions of sequences. The advantage of these methods is that the sequence statistics can be extracted as completely as possible, but it takes several weeks even months and plenty of computing resources to get the embedded models. Fortunately, by using the idea of transfer learning, these models can be used as pretraining models to directly apply to new tasks such as ACP prediction in this study [59].

In the work, we developed a new machine learning method named iACP-DRLF to predict the ACPs from peptide sequences. It was designed to use two kinds of deep representation learning feature extraction technologies to convert the sequences into feature vectors, and used the light gradient boosting machine (LGBM) feature selection to determine the best feature space. After optimizing, iACP-DRLF achieved good 5-fold cross-validation and independent testing accuracy as compared to the previously top two methods, i.e. ACPred-Fuse [25] and AntiCP 2.0 [16]. The two feature analysis approaches, including uniform manifold approximation and projection for dimension reduction (UMAP) [60] and shapley additive explanations (SHAP) [61], were also used to explore the effect of different deep representation learning features on the model performance. Executable and easily-used python scripts are publicly accessible.

Methods and materials

The modeling flowchart is shown in Figure 1. First, the peptide sequences were embedded into feature vectors using two pretrained deep representation learning embedding models (soft symmetric alignment [SSA] and UniRep) to obtain two types of features: SSA features (121D) and UniRep features (1900D). Second, the features were fed into six machine learning models. Third, the selected SSA and UniRep fusion features were used to optimize the six models. Details of the modeling are described in the following sections.

Benchmark dataset

Here, we used the updating benchmark datasets as used in AntiCP 2.0 for modeling and for subsequent comparisons convenient. One dataset is called the main dataset. It contains 861 experimentally validated ACPs and 861 non-ACPs, which was split into two sub-datasets for 5-fold validation training and independent testing. The other dataset was called the alternate dataset consisting of 970 experimentally validated ACPs and 970 non-ACPs, which was also divided into a training subset and independent testing subset. Both datasets could be downloaded from https://webs.iiitd.edu.in/raghava/anticp2/. The ACPs in both datasets were extracted from CancerPPD database [15]. The major difference of the two dataset was that the negative samples of the main dataset were the antimicrobial peptides (AMPs) whereas the negative samples of the alternate dataset were random peptides, which were assumed to be non-ACPs.

Feature extraction

In contrast to the feature extraction methods such as iLearn [62], BioSeq-Analysis [63], Pse-in-One [64] and iFeature [65], two sequence deep representation learning embedding methods were used in this study. The embedding procedure is illustrated in Figure 1. They are available at https://github.com/tbepler/protein-sequence-embedding-iclr2019, https://github.com/churchlab/UniRep. A NVIDIA GPU is required for sequence embedding.

Pretrained SSA embedding

At first, the peptide sequences were fed into a pretrained language model trained on the Pfam dataset. Then the encoded outputs were used as three layers of stacked BiLSTM encoders following a linear layer to get the final embedding matrix $R^{L\times 121}$ for each peptide sequence, where L is the length of the peptide. We called this model SSA embedding because it was trained

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

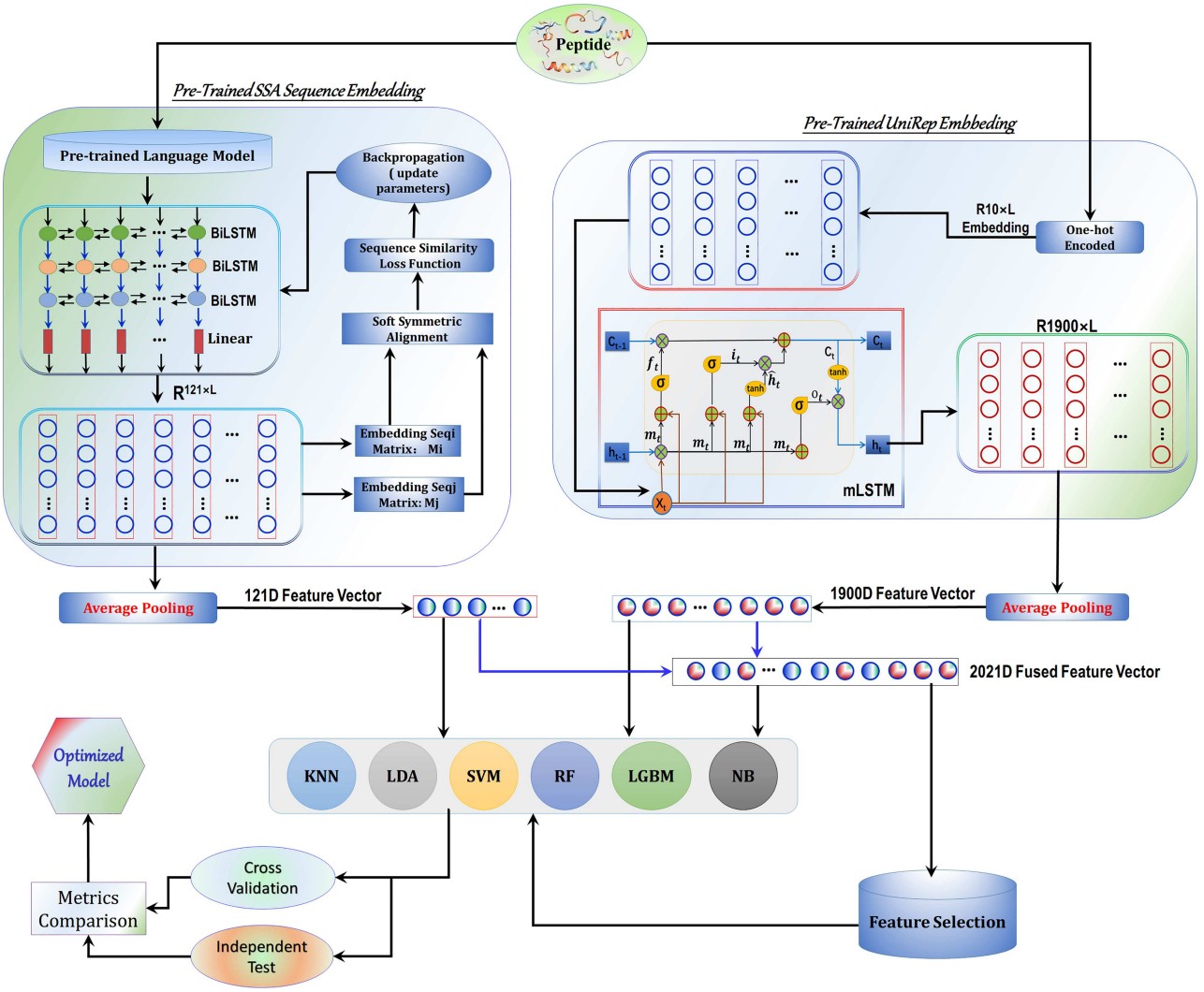

Figure 1. Overview of the modeling. The peptide sequences were first embedded into feature vectors by using the pretrained SSA sequence embedding model and UniRep embedding model and converted into 121 dimension (D) SSA feature vectors and 1900 dimension UniRep feature vector for each sequences. The SSA features, UniRep features and fused SSA-UniRep features (2021D) were then used as input for KNN, LDA, SVM, RF, LGBM and NB predictors. Also, the six models were optimized by feature selection methods. After comparison for cross-validation and independent test metric scores, the optimized model was attained.

and optimized by using a mechanism called SSA. We supposed the presence of two $R^{L\times 121}$ embedded matrices S_1 and S_2 of two different peptide sequences with lengths L_1 and L_2, respectively.

$$S_1 = [x_1, x_2, \cdots, x_{L1}], \tag{1}$$

where x_i is a vector with 121D.

$$S_2 = [y_1, y_2, \cdots,], \tag{2}$$

where y_i is a vector with 121D.

The similarity of S_1 and S_2 was calculated using the Equation (3).

$$\hat{s} = -\frac{1}{A}\sum_{i=1}^{L1}\sum_{j=1}^{L2} a_{ij}\left\|x_i - y_j\right\|_1 \tag{3}$$

with a_{ij}determined using the following Equations (4)–(7).

$$\delta_{ij} = \frac{\exp\left(-\left||x_i - y_k\right||_1\right)}{\sum_{k=1}^{L2}\exp\left(-\left||x_i - y_k\right||_1\right)} \tag{4}$$

$$\varepsilon_{ij} = \frac{\exp\left(-\left||x_k - y_j\right||_1\right)}{\sum_{k=1}^{L1}\exp\left(-\left||x_k - y_j\right||_1\right)} \tag{5}$$

$$a_{ij} = \delta_{ij} + \varepsilon_{ij} - \delta_{ij}\varepsilon_{ij} \tag{6}$$

$$A = \sum_{i=1}^{L1}\sum_{j=1}^{L2} a_{ij} \tag{7}$$

The SSA embedding model was then fit by the above parameters of the sequence encoding via backpropagation with a structure similarity loss function as described in reference [44]. The finally trained model will convert a peptide sequence to an embedded matrix $R^{L\times 121}$. An average pooling operation is used to yield a 121D SSA feature vectors.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

Pretrained UniRep embedding

A sequence with L amino acid residues was first encoded using one-hot encoding, and was then embedded into a $R^{L\times 10}$ matrix. The matrix was fed into an mLSTM encoder as shown in Figure 1 to attain a $R^{1900\times L}$ hidden state output as the embedding matrix. After an average pooling operation, a 1900D vector called a UniRep feature in this work was obtained. The mLSTM encoder calculations involved the use of the Equations (8)–(14).

$$m_t = (X_t W_{xm}) \otimes (h_{t-1} W_{hm}) \tag{8}$$

$$\hat{h}_t = \tan h\,(X_t W_{xh} + m_t W_{mh}) \tag{9}$$

$$f_t = \sigma\,(X_t W_{xf} + m_t W_{mf}) \tag{10}$$

$$i_t = \sigma\,(X_t W_{xi} + m_t W_{mi}) \tag{11}$$

$$o_t = \sigma\,(X_t W_{xo} + m_t W_{mo}) \tag{12}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \hat{h}_t \tag{13}$$

$$h_t = o_t \otimes \tan h\,(C_t) \tag{14}$$

where $\otimes$ indicates element-wise multiplication, h_{t-1} denotes the previous hidden state, X_t is the current input, m_t is the current intermediate multiplicative state, $\hat{h}_t$ is the input before the hidden state, f_t is the forget gate, i_t is the input gate, o_t is the output gate, C_{t-1} is the previous cell state, C_t is the current cell state and h_t is the output hidden state. Also, σ is the sigmoid function and tan$h$ is the hyperbolic tangent function. The UniRep model was pretrained by performing next amino acid prediction with cross-entropy loss minimization.

The fusion of features involved specifically combining 121D SSA features with 1900D features to attain 2021D fused features.

For comparison, we also used non-embedding feature extraction methods. The methods included AAC, DPC, PseAAC and amphiphilic pseudo-amino acid composition (AmPseAAC), which were generated by use of toolkit iFeature (https://github.com/Superzchen/iFeature). The details of these methods are in the reference [65]. Another peptides sequence feature extraction method was also used, namely Word2Vec (W2V). The W2V feature is generated by the biovec toolkit (https://github.com/kyu999/biovec) [43].

Feature selection method

Feature selection is widely used to overcome model overfitting and get the best and the most important feature space for modeling optimization [66–70]. Feature selection methods such as analysis of variance (ANOVA), minimum redundancy maximum relevance (mrmr) and Maximum-Relevance-Maximum-Distance (MRMD) [71, 72] have been proposed. Here, we used LGBM to seek out the best feature space and rank the features in an order according to the feature importance values. The LGBM feature selection has been used for RNA pseudouridine site [33] and DNA methycytosine site predictions [73, 74]. Here is the specific and brief detail about LGBM feature selection. First, input the data and its label into a LGBM model and fit the model. Then with the in-built function in the LGBM model, the importance value for each feature could be obtained. Rank and sort the features from the largest to the smallest according the feature importance values. All features with importance value larger than the critical value (e.g. the average feature importance value) are selected. A code for LGBM feature selection is available at https://github.com/zhibinlv/iACP-DRLF.

Machine learning methods

Six widely used machine learning methods were adopted for comparison. They were K-nearest neighbors (KNN) [75], linear discriminant analysis (LDA) [76], support vector machine (SVM) [73, 77–81], RF [82–87], LGBM [88] and naive Bayes (NB) [89].

Evaluation metrics and methods

To evaluate the model performance, we used total accuracy (ACC), sensitivity (Sn), specificity (Sp) and Matthews correlation coefficient (MCC) and area under receiver operating characteristic curves (AUC). The given true positive sample number (TP), true negative sample number (TN), false positive sample number (FP) and false negative sample number (FN) were used to compute the metrics using the equations [90–100].

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{15}$$

$$\text{Sn} = \frac{TP}{(TP + FN)} \tag{16}$$

$$\text{Sp} = \frac{TN}{TN + FN} \tag{17}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{18}$$

The AUC is defined as the area under the receiver operating characteristic curve (ROC). The ROC is drawn according to a series of different cut-off values or thresholds, with true positive rate as the ordinate and false positive rate as the abscissa.

K-fold cross-validation and independent testing are widely used methods to evaluate machine learning model [53, 101–104]. K-fold cross-validation divides the original data into K groups (K-fold), conducts a validation for each data subset and uses the remaining K-1 data subsets as the training set. These K models are evaluated in the validation set respectively. The final values of the metrics of the models are averaged to obtain the cross-validated values. The 5-fold (K = 5) cross-validation method was used in the current work. When carrying out independent testing, a dataset completely different from the training dataset is used. That is, all the samples are new to the trained model.

Results and discussion

Initial performance of models trained on the main dataset

To find out the better embedding feature types, we first developed models based on six machine learning methods using SSA and UniRep embedding features. The 5-fold cross-validation scores of different models utilized different features are shown in Figure 2. The average independent testing scores of these models are listed in Supplementary Table S1. For 5-fold cross-validation accuracy, it could be observed from Figure 2 that models except for LDA based on UniRep features were with better performance than models based on SSA features. For example, the average ACC of KNN, SVM, RF, LGBM and NB with UniRep features were 72.3, 75.5, 72.6, 74.3 and 64.8%, respectively, which were over to models with SSA features by value of 3.64, 14.1, 4.38, 3.85 and 3.23%. Although the ACC value 61.5% of LDA model with UniRep features was lower than that of model with SSA feature by value of 6.94%. The UniRep features were better than the SSA features for the ACPs and the non-ACPs identification, which were further confirmed by the UMAP feature visualization

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

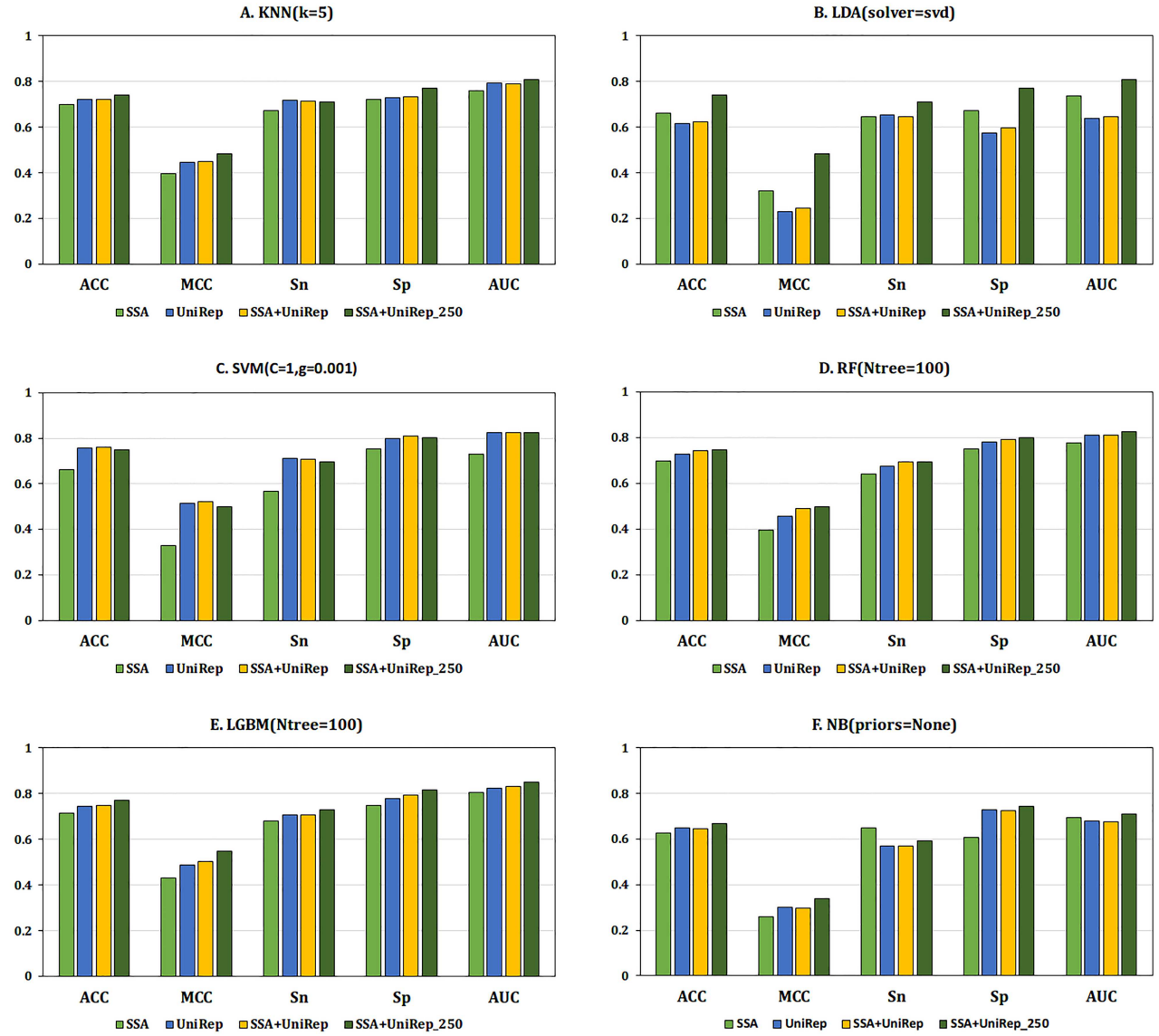

Figure 2. Five-fold cross-validation metrics comparison for six models based on SAA, UniRep, fused SSA + UniRep features and the top 250 selected SSA + UniRep fused features (SSA + UniRep_250). The trained dataset was the main dataset. The six models were KNN, LDA, SVM, RF, LGBM and NB. The light green bars are for models based on the SSA features, the blue bars are models based on the UniRep features, the yellow bars are for models based on the fused SSA + UniRep and the dark green bars are for models based on the SSA + UniRep_250 features.

technology as shown in Figure 5A and B. Clearly, after UMAP feature reduction, most of the APCs samples represented by UniRep could be apart from non-ACPs samples easily as displayed in Figure 5B, whereas the ACPs and non-ACPs represented by SSA were still all mixed up as shown in Figure 5A.

For further improving the model performance, we used the feature fusion strategy and feature selection technology. We fused the SSA features (121D) with the UniRep features (1900D) to yield the SSA + UniRep features (2021D). Then the 2021D SSA + UniRep features were fed into a LGBM model to compute the feature importance values. According to descending sort feature importance, the top 250 features (SSA + UniRep_250) were selected for modeling. It meant the NO.1 feature was with the largest feature importance value. The 5-fold cross-validation scores of models used SSA + UniRep features and SSA + UniRep_250 features are displayed in Figure 2 and the values are listed in Supplementary Table S1. As we could see from Figure 2 and Supplementary Table S2, the accuracy of models except for RF and NB with SSA + UniRep features were improved no >0.8%, whereas the ACC of RF increased by 1.6% and the ACC of NB decreased slightly by 0.14%. That was, fusing the two features type directly could improve some models' performance, but it did not always work. The UMAP results shown in Figure 5A–C confirmed that if the SSA features fusing with the UniRep features, the ACPs and the non-ACPs would be confounded again; but if the feature selection technology was used, the ACPs and non-ACPs could be separated as displayed in the UMAP figures (Figure 5D).

As compared with models with SSA and UniRep features, the accuracy of models except for SVM with the selected SSA + UniRep_250 features were with great improvement by 2.5–19.8%. Obviously, the ACC of the LDA model with SSA + UniRep_250 increased dramatically by 19.8% over that without feature selection. It could be seen from Figures 2

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

and 4 and Supplementary Table S2 that the LGBM model using deep representation learning with SSA + UniRep_250 features were with the best 5-fold cross-validation score (ACC = 77.2%, MCC = 0.547, Sp = 81.5% and AUC = 0.851) except for Sn = 72.8% and with the best average independent testing scores (ACC = 75.4%, MCC = 0.509, Sn = 78.3%, Sp = 72.5% and AUC = 0.817).

The advantage of SHAP value is that the SHAP mean values reflect the impact of the feature on the sample identification or the contribution of the feature to the sample identification; the larger the SHAP value is the more impact or contribution of the feature to the sample identification [61]. As the SHAP feature importance analysis shown in Figure 4C, for LGBM model, the 20 features with the top SHAP values included 18 UniRep features and 2 SSA features. In terms of feature quantity, more UniRep features than SSA features contributed to the ACPs and the non-ACPs identification. In terms of feature SHAP values, the UniRep_F1411 (the 1411st dimension UniRep feature) was with larger mean SHAP value than the SSA_F111. It means that UniRep_F1411 was with more impact on the ACPs and non-ACP identification. The UniRep_F1411 could be more easily to discriminate the ACPs from the non-ACPs than SSA_F111. Also, the comparison between UniRep_F270 and SSA_F91 was similar. Both from the quantity and quality of the features list in Figure 4C, the UniRep features played more important role than the SSA features for ACPs identification.

Initial performance of models trained on the alternate dataset

To further explore the advantages of peptides sequences deep representation learning features, we trained models on the alternate dataset and used the models to identify the ACPs from the random non-ACPs.

The 5-fold cross-validation scores of six models are shown in Figures 3 and 4. The exact value of the cross-validation scores and independent testing scores are listed in Supplementary Table S2. It can be observed that models except for LDA with UniRep features were better than those with SSA features. As compared with models with SSA, UniRep and SSA + UniRep feature, the cross-validation scores of AAC, MCC, Sn and AUC of model with SSA + UniRep_250 feature increased obviously. It means that LGBM feature selection technology effectively enhanced the models performance, especially for KNN, NB and LDA models (Figure 4). Similar to the main dataset, models with SSA + UniRep_250 were with better performance for almost all cases. The LGBM based model developed for alternate dataset using SSA + UniRep_250 was with the best cross-validation scores (ACC = 92.9%, MCC = 0.859, Sn = 91.7% and AUC = 0.980) except for Sn = 94.1%.

In the case of independent testing scores shown in Supplementary Table S2, the LDA model based on UniRep was with the poorest performance while models based SSA + UniRep_250 were with very close scores for ACC ranging from 89.9 to 92.1%, MCC ranging from 0.791 to 0.944, Sn ranging from 87.6 to 89.1% and Sp ranging from 90.9 to 96.2%. Unlike the SSA + UniRep_250 feature based LGBM model trained on the main dataset, the LGBM model trained on the alternate dataset was not the top model with the best independent test scores, although it has fairly good test performance (ACC = 91.7%), slightly lower than the value 92.2% of RF model with SSA feature. To find out the best model has been done by hyper-parameter searching optimization and it would be discussed in another following section.

The UMAP feature visualization for the alternate dataset is shown in Figure 5. It shared the same commons and trends as that for the main dataset. And the SHAP feature importance analysis for the alternate dataset are shown in Figure 4D. The top 20 important feature for alternate dataset consisted of 19 UniRep features and 1 SSA feature, which came in twentieth. Similar to the analysis in Section Initial performance of models trained on the main dataset, the results proved that the UniRep has an advantage over SSA in identifying the ACPs from the non-ACPs for the alternate dataset.

Comparison models with different feature types

To explore DRLF and non-DRLF features effect on ACPs identification, four classical peptide sequences feature extraction technologies (AAC, DPC, PseAAC and AmPseAAC) and one non-deep learning embedding technology (W2V) were used. The ACC and MCC values of the developed models based on the above features are shown in Tables 1, 2, 3 and 4. The best value of ACC and MCC for every model based on different features are underline and in bold. More metrics values are listed in Supplementary Tables S1 and S2.

In the case of models trained on the main dataset, the cross-validation ACC of models (except for SVM and RF) using SSA + UniRep_250 features were over those of models using other features. Although the cross-validation MCC values of RF used DPC and NB used AAC were better than LGBM used SSA + UniRep_250, their ACC and MCC values were inferior to the values of LGBM. For independent test on the main dataset, the LGBM using SSA + UniRep_250 obtained the best ACC (75.4%) and MCC (0.510).

In the case of models trained on the alternate dataset, all the models using SSA + UniRep_250 were with the best cross-validation ACC and MCC as compared with models using other feature. For independent testing on the alternate dataset, the models (except for RF and LGBM using) SSA + UniRep_250 had better accuracy. The RF using AmPseAAC was with good ACC (91.7%) and MCC (0.835), but it was inferior to the values (ACC = 92.1% and MCC = 0.844) of SVM model using SSA + UniRep_250 and the LGBM model using UniRep.

Evidently, for most cases, the SSA + UniRep_250 feature was with advantages over using other features for modeling; the models based on DRLF features archived better cross-validation accuracy and independent testing accuracy.

Model optimization and prediction script implementation

To determine the optimal model for the ACPs prediction, we used the feature increment strategy and the hyper-parameter grid searching method to obtain the best model using SSA + UniRep_250 features. The feature increment strategy was to construct 250 models by using the top 1, 2, ..., 250 features. For each model, its hyper-parameter searching was to use the scikit-learn GridSearchCV module. The best KNN, LDA, SVM, RF, LGBM and NB model trained and tested on the main dataset and the alternate dataset are listed in Tables 5 and 6.

In the case of the main dataset, the LGBM using the top 148 selected SSA + UniRep was with the best 5-fold cross-validation scores (ACC = 79.1%, MCC = 0.583, Sn = 77.0%, Sp = 81.2% and AUC = 0.873) and the best independent testing scores (ACC = 77.4%, MCC = 0.551, Sn = 80.7% and Sp = 74.3%). In the case of models trained and tested on the alternate dataset, the LGBM using the top 129 SSA + UniRep feature

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

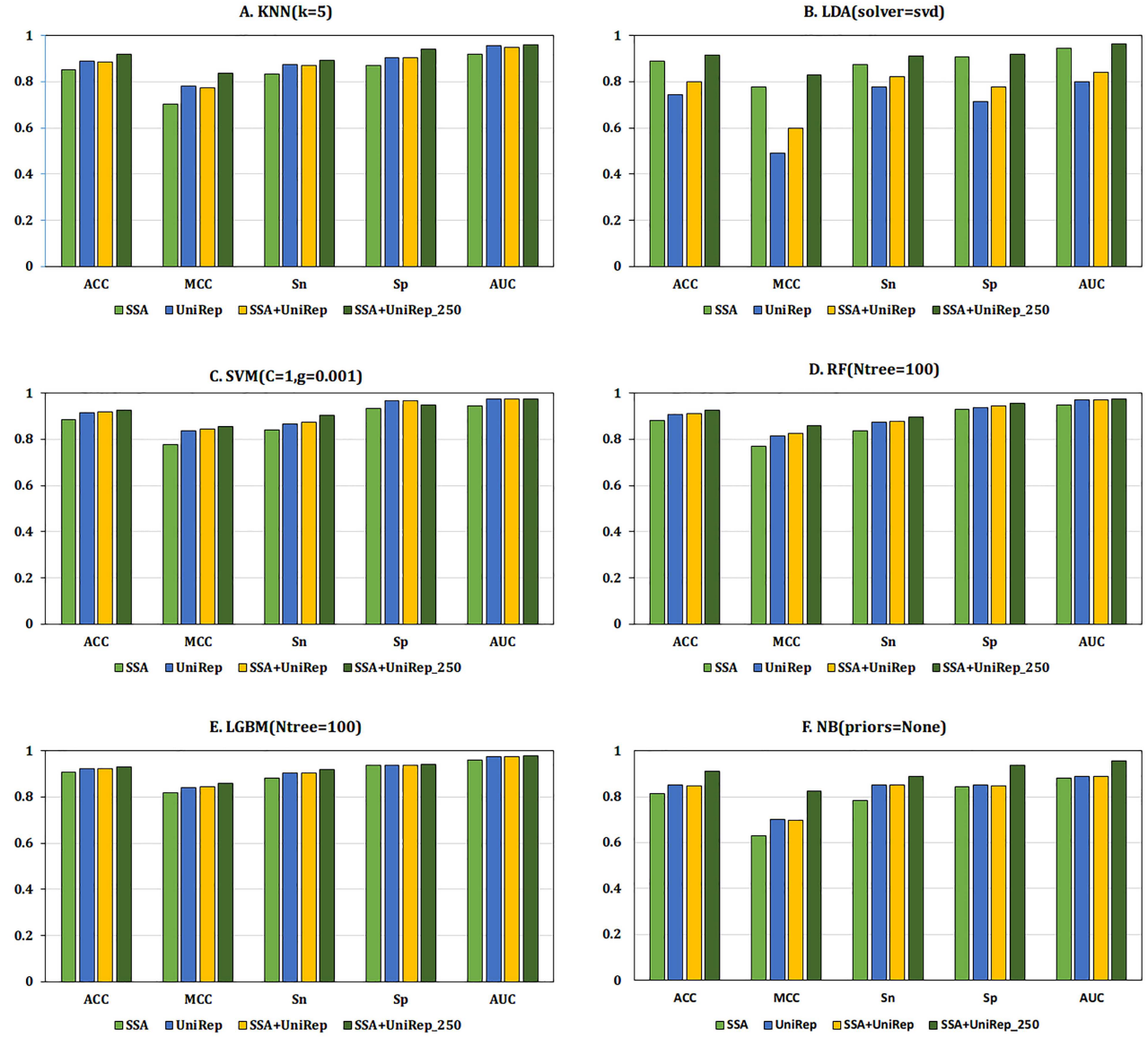

Figure 3. Five-fold cross-validation metrics comparison for six models based on SAA, UniRep, fused SSA + UniRep features and the top 250 selected SSA + UniRep fused features(SSA + UniRep_250). The trained dataset is the alternate dataset. The six models were KNN, LDA, SVM, RF, LGBM and NB. The light green bars are for models based on the SSA features, the blue bars are models based on the UniRep features, the yellow bars are for models based on the fused SSA + UniRep and the dark green bars are for models based on the SSA + UniRep_250 features.

Table 1. Five-fold cross-validation accuracy and sensitivity comparison for six machine learning models with different feature types based on the main trained dataset

Model/metrics	KNN		LDA		SVM		RF		LGBM		NB	
Features type	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
AAC	71.8	0.439	64.2	0.286	64.8	0.321	74.6	0.498	73.3	0.466	66.6	**0.340**
DPC	61.2	0.247	70.2	0.404	73.3	0.471	**75.9**	**0.521**	73.3	0.467	66.6	0.333
PseAAC	70.9	0.418	64.1	0.283	63.4	0.274	74.1	0.486	71.7	0.434	62.1	0.251
AmPseAAC	71.6	0.434	65.1	0.304	64.8	0.300	73.7	0.479	74.1	0.482	62.6	0.261
W2V	71.1	0.424	64.0	0.280	66.4	0.328	69.3	0.387	70.9	0.420	61.1	0.231
SSA	69.8	0.396	66.1	0.323	66.1	0.329	69.6	0.395	71.6	0.433	62.8	0.257
UniRep	72.3	0.448	61.5	0.231	75.5	0.513	72.7	0.456	74.3	0.489	64.8	0.301
SSA + UniRep	72.4	0.448	62.2	0.245	**75.9**	**0.521**	74.3	0.490	75.0	0.502	64.7	0.298
SSA + UniRep_250	**74.1**	**0.485**	**73.7**	**0.475**	74.8	0.499	74.9	0.501	**77.2**	**0.547**	**66.8**	**0.340**

[a]The best value of each column is underline and in bold; ACC unit: percentage(%).

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

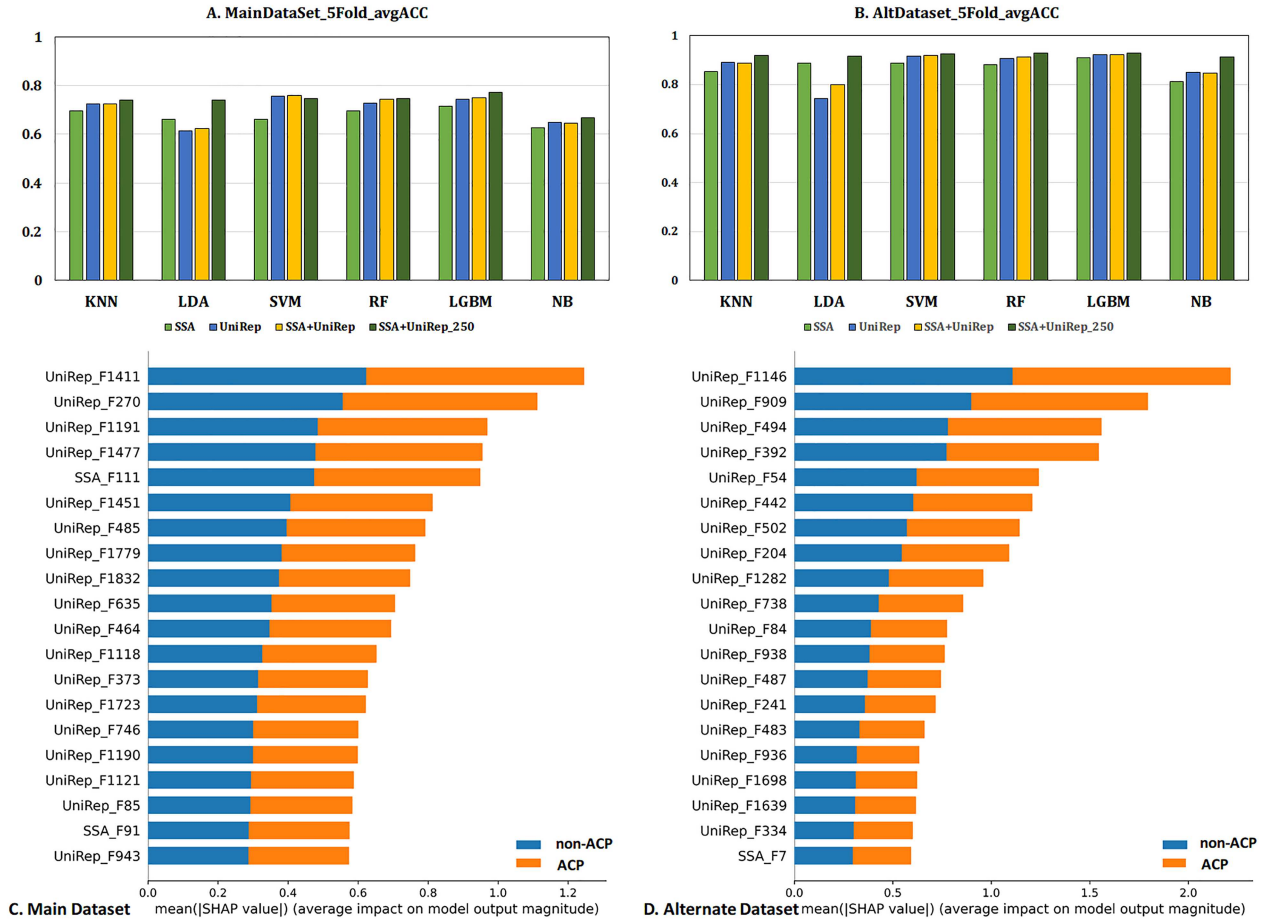

Figure 4. (**A** and **B**) Comparison of the 5-fold average accuracy of KNN, LDA, SVM, RF, LGMB and NB models used SSA, UniRep, SSA + UniRep and SSA + UniRep_250 features. (**C** and **D**) Feature importance analysis using the SHAP method. The blue and orange bars mean the feature impact or contribution for identifying the non-ACPs and ACPs; the lager the bars' length, the more important or the more contribution of the feature for the non-ACPs and ACPs identification.

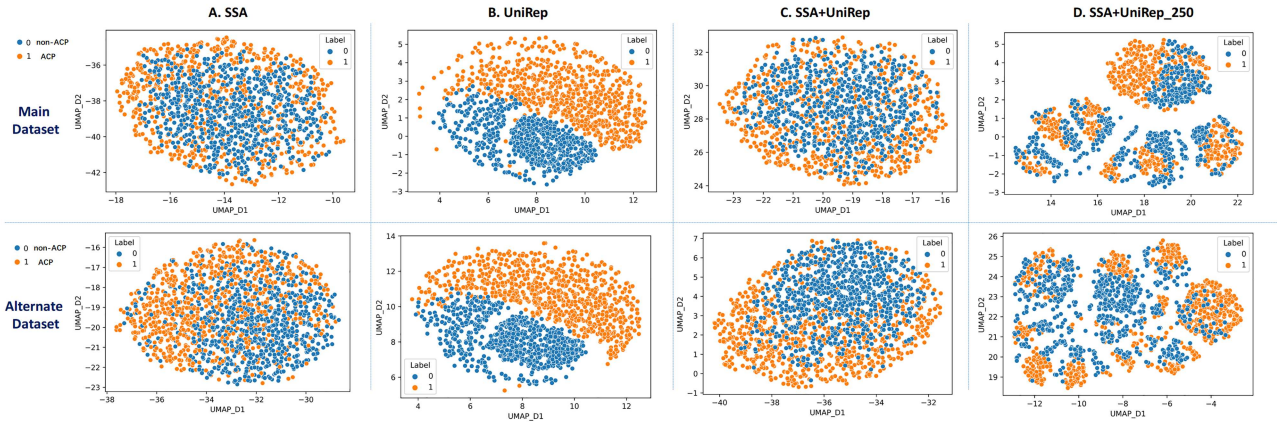

Figure 5. Feature visualization by UMAP for dimension reduction. (**A**) is for the SSA features, (**B**) is for the UniRep features, (**C**) is for the SSA fused UniRep features and (**D**) is for the top 250 features selected from SSA fused UniRep features.

was with ACC = 94.5%, MCC = 0.891, Sn = 92.6%, Sp = 96.4% and AUC = 0.984, which was superior to other models for the cross-validation scores. But the LGBM model independent testing scores (ACC = 93.0%, MCC = 0.862, Sn = 89.6% and Sp = 96.4%) were inferior to the LDA model with the best test scores (ACC = 93.5%, MCC = 0.872%, Sn = 90.2% and Sp = 96.9%). Considering the robustness of the model in practical application, we intended to choose models with better cross-validation accuracy. In addition, we used RF feature selection method to optimize the models. The results are shown in Supplementary Table S3. On the whole view, the LGBM feature selection based methods have superior cross-validation performance over the RF feature selection

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

Table 2. Independent test accuracy and sensitivity comparison for six machine learning models with different feature types based on the main independent testing dataset

Model/metrics	KNN		LDA		SVM		RF		LGBM		NB	
Features type	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
AAC	**70.9**	**0.419**	66.3	0.327	63.8	0.294	73.0	0.460	72.5	0.450	**67.6**	**0.358**
DPC	63.7	0.297	69.1	0.383	71.5	0.430	72.8	0.456	69.8	0.397	70.3	0.408
PseAAC	69.2	0.389	65.4	0.309	64.3	0.290	74.0	0.481	73.2	0.465	66.1	0.334
AmPseAAC	68.8	0.383	64.4	0.290	65.0	0.305	**75.0**	**0.502**	75.0	0.502	66.5	0.342
W2V	68.8	0.382	63.0	0.260	64.7	0.294	67.9	0.358	69.8	0.397	63.3	0.273
SSA	68.0	0.361	68.4	0.367	65.8	0.320	67.0	0.339	69.1	0.383	64.0	0.282
UniRep	67.0	0.340	63.0	0.262	71.9	0.439	71.2	0.425	73.6	0.473	60.8	0.218
SSA + UniRep	68.0	0.360	63.3	0.267	**72.3**	**0.446**	69.9	0.399	72.3	0.448	61.2	0.226
SSA + UniRep_250	70.7	0.414	**69.7**	**0.397**	72.0	0.441	70.9	0.418	**75.4**	**0.510**	62.9	0.260

[a]The best value of each column is underline and in bold; ACC unit: percentage (%).

Table 3. Five-fold cross-validation accuracy and sensitivity comparison for six machine learning models with different feature types based on the alternate trained dataset

Model/metrics	KNN		LDA		SVM		RF		LGBM		NB	
Features type	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
AAC	88.3	0.769	86.7	0.736	86.5	0.733	88.8	0.778	90.7	0.815	86.3	0.729
DPC	64.1	0.344	83.6	0.673	88.1	0.765	87.3	0.746	87.5	0.752	84.5	0.690
PseAAC	88.5	0.771	86.5	0.731	85.0	0.704	89.0	0.782	89.7	0.796	82.4	0.660
AmPseAAC	88.8	0.777	87.2	0.745	84.6	0.698	89.4	0.789	90.2	0.804	82.1	0.653
W2V	79.4	0.592	78.5	0.571	84.6	0.694	84.8	0.699	84.5	0.690	67.4	0.361
SSA	85.2	0.704	88.9	0.779	88.6	0.776	88.3	0.769	90.8	0.818	81.4	0.629
UniRep	89.0	0.780	74.4	0.489	91.6	0.837	90.6	0.815	92.1	0.842	85.0	0.701
SSA + UniRep	88.6	0.744	79.9	0.598	92.0	0.844	91.2	0.825	92.1	0.844	84.8	0.696
SSA + UniRep_250	**91.8**	**0.837**	**91.4**	**0.829**	**92.6**	**0.854**	**92.8**	**0.857**	**92.9**	**0.859**	**91.2**	**0.826**

[a]The best value of each column is underline and in bold; ACC unit: percentage (%).

Table 4. Independent test accuracy and sensitivity comparison for six machine learning models with different feature types based on the alternate independent testing dataset

Model/metric	KNN		LDA		SVM		RF		LGBM		NB	
Features type	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
AAC	90.9	0.820	89.0	0.782	90.1	0.805	91.6	0.832	90.5	0.809	87.3	0.748
DPC	64.5	0.351	83.9	0.680	90.2	0.807	87.6	0.751	86.7	0.735	86.7	0.736
PseAAC	90.7	0.814	86.8	0.736	86.3	0.729	91.2	0.826	91.6	0.833	80.9	0.628
AmPseAAC	90.1	0.802	88.4	0.768	86.0	0.723	**91.7**	**0.835**	91.2	0.824	82.6	0.661
W2V	80.6	0.615	79.5	0.592	86.1	0.722	86.9	0.738	88.2	0.765	64.3	0.294
SSA	88.1	0.762	89.4	0.790	90.9	0.821	92.2	0.848	91.6	0.832	84.3	0.688
UniRep	89.5	0.792	75.6	0.513	91.7	0.839	90.9	0.821	**92.1**	**0.844**	86.7	0.734
SSA + UniRep	89.8	0.798	80.7	0.615	91.7	0.837	90.6	0.815	91.3	0.828	87.4	0.747
SSA + UniRep_250	**91.7**	**0.836**	**91.1**	**0.824**	**92.1**	**0.844**	91.2	0.826	91.7	0.835	**89.5**	**0.791**

[a]The best value of each column is underline and in bold; ACC unit: percentage(%).

based methods. Thus, the LGBM models based on LGBM feature selection were determined to be the finally used models in the prediction script implementation.

The python source code of our method named iACP-DRLF and the pertained models are available at https://github.com/zhibinlv/iACP-DRLF. Our iACP-DRLF is a user-friendly and easily used method. It runs with a FASTA format file as input and it outputs the results in a CSV file. However, it requires a NVIDIA GPU to accelerate the computation. We hope that iACP-DRLF would become a useful tool for reader interest in ACPs prediction.

Comparison with the existing methods

We also compared our iACP-DRLF method to the existing methods for independent testing. The independent testing scores for state-of-the-art ACP prediction came from references [16]. The results are shown in Table 7. It could be observed that the ACC (77.5 and 93.0%) and MCC (0.55 and 0.86) of iACP-DRLF outperformed the reported existing methods on the both dataset. For the main dataset, all methods except for PEPred-Suite were with higher sensitivity than specificity. The good performance of iACP-DRLF tested on the main dataset was due to

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

Table 5. The optimized models (trained on the main dataset) metrics comparison

Model (parameters)	Feature Dims	5-Fold cross-validation					Independent testing			
		ACC	MCC	Sn	Sp	AUC	ACC	MCC	Sn	Sp
KNN (k = 5)	33	74.0%	0.480	72.4%	75.6%	0.797	73.7%	0.475	77.8%	69.6%
LDA (solver = 'svd)	240	73.3%	0.465	73.5%	73.0%	0.816	73.1%	0.465	78.9%	67.3%
SVM (C = $10^{0.8}$, g = $10^{-2.4}$)	235	77.8%	0.557	76.9%	78.8%	0.853	76.3%	0.530	**81.9%**	70.8%
RF (Ntree = 250,Nleaf = 2)	211	76.7%	0.536	72.2%	81.1%	0.841	73.4%	0.468	76.0%	70.8%
LGBM (Ntree = 400,depth = 12)	148	**79.1%**	**0.583**	**77.0%**	**81.2%**	**0.873**	**77.5%**	**0.551**	80.7%	**74.3%**
NB (priors = None)	16	67.4%	0.349	65.3%	69.6%	0.743	66.7%	0.334	63.2%	70.2%

[a]The best value of each column is underline and in bold; (Dims: dimensions).

Table 6. The optimized models (trained on the alternate dataset) metrics comparison

Model (parameters)	Feature Dims	5-Fold cross-validation					Independent testing			
		ACC	MCC	Sn	Sp	AUC	ACC	MCC	Sn	Sp
KNN (k = 5)	220	92.8%	0.857	91.2%	94.5%	0.965	92.7%	0.857	89.6%	95.9%
LDA (solver = svd)	200	91.5%	0.830	90.3%	92.6%	0.964	**93.5%**	**0.872**	**90.2%**	**96.9%**
SVM (C = $10^{0.27}$, g = $10^{-1.87}$)	123	93.8%	0.877	91.4%	96.3%	0.979	92.0%	0.843	87.6%	96.4%
RF (Ntree = 150,Nleaf = 2)	91	93.0%	0.861	89.7%	96.3%	0.978	92.5%	0.852	88.6%	96.4%
LGBM (Ntree = 250, depth = 9)	129	**94.5%**	**0.891**	**92.6%**	**96.4%**	**0.984**	93.0%	0.862	89.6%	96.4%
NB (priors = None)	86	90.9%	0.820	87.7%	94.1%	0.956	91.5%	0.832	87.6%	95.3%

[a]The best value of each column is underline and in bold; (Dims: dimensions).

Table 7. Comparison of the independent testing metrics values for iACP-DRLF with state-of-the-art ACP predictors

Methods	Main dataset				Alternate dataset			
	ACC	MCC	Sn	Sp	ACC	MCC	Sn	Sp
iACP-DRLF	**77.5%**	**0.55**	80.7%	**74.3%**	**93.0%**	**0.86**	89.6%	**96.4%**
AntiCP_2.0	75.4%	0.51	77.5%	73.4%	92.0%	0.84	**92.3%**	91.8%
AntiCP	50.6%	0.07	**100.0%**	1.2%	90.0%	0.80	89.7%	90.2%
ACPred	53.5%	0.09	85.6%	21.4%	85.3%	0.71	87.1%	83.5%
ACPred-FL	44.8%	−0.12	67.1%	22.5%	43.8%	−0.15	60.2%	25.6%
ACPpred-Fuse	68.9%	0.38	69.2%	68.6%	78.9%	0.60	64.4%	93.3%
PEPred-Suite	53.5%	0.08	33.1%	73.8%	57.5%	0.16	40.2%	74.7%
iACP	55.1%	0.11	77.9%	32.2%	77.6%	0.55	78.4%	76.8%

[a]The best value of each column is underline and in bold.

its balanced sensitivity and specificity. For the alternate dataset, the Sn (89.7%) of iACP-DRLF ranked 3rd, but its Sp (96.4%) ranked top as shown in Table 7. The better specificity contributed more than sensitivity to the good performance of iACP-DRLF as compared to other methods. All the observation indicates that iACP-DRLF is one of the machine learning based state-of-the-art ACP predictors. In contrast to the non-DRLF used by other methods, the DRLF used by iACP-DRLF was able to distinguish the ACPs from the non-ACPs more accurately.

Conclusions

Overall, we have developed a new method named iACP-DRLF, involving the use of a sequence-based deep representation learning feature embedding method to predict potential ACPs. This was the first time, to the best of our knowledge, that the deep representation learning features were adopted for ACPs prediction. By carrying out feature fusion and feature selection, we obtained two optimal LGBM models for two datasets respectively. Meanwhile, using the same training and testing benchmark datasets, iACP-DRLF yielded ACC and MCC exceeding those of the previous predictive models for the same kind of task. With the use of deep representation learning features, iACP-DRLF demonstrated a better ability to identify ACPs from non-ACPs. The UMAP feature visualization and the SHAP value feature importance analysis showed that UniRep features played a more important role than SSA features for ACPs prediction. The pretrained models and a user-friendly python script for iACP-DRLF were also publicly available for readers.

Although the use of deep representation learning features improved model prediction performance, the specific physical meanings of these features are unclear. Also in order to obtain these deep representation learning features quickly, a GPU-accelerated computation resource is usually required. However, these drawbacks do not keep us from continuing to apply this method to peptide or protein sequence analysis tasks, such

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

as protein subcellular localization, protein post-transcriptional modification, prediction of signal peptides, artificial protein design, etc. [105–110].

Key Points

- A new method called iACP-DRLF for predicting ACP with high accuracy was developed.
- Unlike previously methods, iACP-DRLF used deep representation learning feature embedding technology.
- The performance of iACP-DRLF was superior to the methods used non-DRLF.
- The UMAP feature visualization and SHAP value analysis proved that the UniRep features were better features for ACPs prediction.
- The method iACP-DRLF was available as python script.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

The project is funded by the National Natural Science Foundation of China (Nos. 62001090, 61922020, 61771331), and by the China Postdoctoral Science Foundation (No. 2020M673184).

References

1. Bray F, Ferlay J, Soerjomataram I, *et al*. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries (vol 68, pg 394, 2018). *Ca-a Cancer J Clin* 2018. doi: 10.3322/caac.21609:1.
2. Cheng L, Hu Y. Human disease system biology. *Curr Gene Ther* 2018;**18**:255.
3. Cheng L, Hu Y, Sun J, *et al*. Dincrna: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncrna function. *Bioinformatics* 2018;**34**:1953.
4. Morel D, Jeffery D, Aspeslagh S, *et al*. Combining epigenetic drugs with other therapies for solid tumours - past lessons and future promise. *Nat Rev Clin Oncol* 2020;**17**:91.
5. Achrol AS, Rennert RC, Anders C, *et al*. Brain metastases. *Nat Rev Dis Primers* 2019;**5**:26.
6. Cheng L. Computational and biological methods for gene therapy. *Curr Gene Ther* 2019;**19**:210.
7. Cheng L, Zhao H, Wang P, *et al*. Computational methods for identifying similar diseases. *Mol TherNucl Acids* 2019;**18**:590.
8. Thakkar S, Sharma D, Kalia K, *et al*. Tumor microenvironment targeted nanotherapeutics for cancer therapy and diagnosis: a review. *Acta Biomater* 2020;**101**:43.
9. Maeda H, Khatami M. Analyses of repeated failures in cancer therapy for solid tumors: poor tumor-selective drug delivery, low therapeutic efficacy and unsustainable costs. *Clin Transl Med* 2018;**7**:20.
10. Chiangjong W, Chutipongtanate S, Hongeng S. Anticancer peptide: physicochemical property, functional aspect and trend in clinical application (review). *Int J Oncol* 2020;**57**:678.
11. Ge RQ, Feng GW, Jing XY, *et al*. Enacp: an ensemble learning model for identification of anticancer peptides. *Front Genet* 2020;**11**:12.
12. Soon TN, Chia AYY, Yap WH, *et al*. Anticancer mechanisms of bioactive peptides. *Protein Pept Lett* 2020. doi: 10.2174/0929866527666200409102747.
13. Dissanayake S, Denny WA, Gamage S, *et al*. Recent developments in anticancer drug delivery using cell penetrating and tumor targeting peptides. *J Control Release* 2017;**250**:62.
14. Pelliccia S, Amato J, Capasso D, *et al*. Bio-inspired dual-selective bcl-2/c-myc g-quadruplex binders: design, synthesis, and anticancer activity of drug-like imidazo 2,1-i purine derivatives. *J Med Chem* 2020;**63**:2035.
15. Tyagi A, Tuknait A, Anand P, *et al*. Cancerppd: a database of anticancer peptides and proteins. *Nucleic Acids Res* 2015;**43**:D837.
16. Agrawal P, Bhagat D, Mahalwal M, *et al*. Anticp 2.0: an updated model for predicting anticancer peptides. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa153.
17. Boopathi V, Subramaniyam S, Malik A, *et al*. Macppred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int J Mol Sci* 2019;**20**:31013619.
18. Tyagi A, Kapoor P, Kumar R, *et al*. In silico models for designing and discovering novel anticancer peptides. *Sci Rep* 2013;**3**:8.
19. Hajisharifi Z, Piryaiee M, Beigi MM, *et al*. Predicting anticancer peptides with chou's pseudo amino acid composition and investigating their mutagenicity via ames test. *J Theor Biol* 2014;**341**:34.
20. Vijayakumar S, Lakshmi PTV. Acpp: a web server for prediction and design of anti-cancer peptides. *Int J Pept Res Ther* 2015;**21**:99.
21. Chen W, Ding H, Feng P, *et al*. Iacp: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 2016;**7**:16895.
22. Zhang JR, Ju Y, Lu HJ, *et al*. Accurate identification of cancerlectins through hybrid machine learning technology. *Int J Genomics* 2016. doi: 10.1155/2016/7604641:11.
23. Manavalan B, Basith S, Shin TH, *et al*. Mlacp: machine-learning-based prediction of anticancer peptides. *Oncotarget* 2017;**8**:77121.
24. Wei LY, Zhou C, Chen HR, *et al*. Acpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 2018;**34**:4007.
25. Rao B, Zhou C, Zhang G, *et al*. Acpred-fuse: fusing multi-view information improves the prediction of anticancer peptides. *Brief Bioinform* 2020;**21**:1846.
26. Wei LY, Zhou C, Su R, *et al*. Pepred-suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics* 2019;**35**:4272.
27. Yi H-C, You Z-H, Zhou X, *et al*. Acp-dl: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Mol Ther - Nucl Acids* 2019;**17**:1.
28. Basith S, Manavalan B, Shin TH, *et al*. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med Res Rev* 2020;**40**:1276.
29. Singh M, Kumar V, Sikka K, *et al*. Computational design of biologically active anticancer peptides and their interactions with heterogeneous popc/pops lipid membranes. *J Chem Inf Model* 2020;**60**:332.
30. Chung CR, Kuo TR, Wu LC, *et al*. Characterization and identification of antimicrobial peptides with different functional activities. *Brief Bioinform* 2020;**21**:1098.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

31. Lv Z, Ao C, Zou Q. Protein function prediction: from traditional classifier to deep learning. *Proteomics* 2019;**19**:1900119.
32. Jin S, Zeng X, Xia F, *et al*. Application of deep learning methods in biological networks. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa043.
33. Lv Z, Zhang J, Ding H, *et al*. Rf-pseu: a random forest predictor for rna pseudouridine sites. *Front Bioeng Biotechnol* 2020;**8**:134.
34. Young T, Hazarika D, Poria S, *et al*. Recent trends in deep learning based natural language processing. *IEEE Comput Intell Mag* 2018;**13**:55.
35. Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks Ananthan Nambiar, Simon Liu, Mark Hopkins, Maeve Heflin, View ORCID ProfileSergei Maslov, Anna Ritz. doi: https://doi.org/10.1101/2020.06.15.153643.
36. Alley EC, Khimulya G, Biswas S, *et al*. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**:1315.
37. Liu B, Gao X, Zhang H. Bioseq-analysis2.0: an updated platform for analyzing DNA, rna, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res* 2019;**47**: e127.
38. Hong Z, Zeng X, Wei L, *et al*. Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 2020;**36**:1037.
39. Zou Q, Chen L, Huang T, *et al*. Machine learning and graph analytics in computational biomedicine. *Artif Intell Med* 2017;**83**:1–1. doi: 10.1016/j.artmed.2017.09.
40. Xu Y, Wang Y, Luo J, *et al*. Deep learning of the splicing (epi) genetic code reveals a novel candidate mechanism linking histone modifications to esc fate decision. *Nucleic Acids Res* 2017;**45**:12100.
41. Junwei H, Xudong H, Qingfei K, *et al*. Pssubpathway: a software package for flexible identification of phenotype-specific subpathways in cancer progression. *Bioinformatics* 2020;**36**:2303.
42. Zhao T, Hu Y, Peng J, *et al*. Deeplgp: a novel deep learning method for prioritizing lncrna target genes. *Bioinformatics* 2020;**36**:4466.
43. Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* 2015;**10**:e0141287.
44. Bepler T, Berger B. Learning protein sequence embeddings using information from structure. 2019; arXiv:1902.08661.
45. Nambiar A, Liu S, Hopkins M, *et al*. Transforming the language of life: transformer neural networks for protein prediction tasks. *BioRxiv* 2020. doi: 10.1101/2020.06.15.153643.
46. Rao R, Bhattacharya N, Thomas N, *et al*. Evaluating protein transfer learning with tape. 2019; arXiv:1906.08230.
47. Strodthoff N, Wagner P, Wenzel M, *et al*. Udsmprot: universal deep sequence models for protein classification. *Bioinformatics* 2020;**36**:2401.
48. Ofer D, Linial M. Profet: feature engineering captures high-level protein functions. *Bioinformatics* 2015;**31**: 3429.
49. Liu B, Li C, Yan K. Deepsvm-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief Bioinform* 2020;**21**:1733.
50. Wei L, Ding Y, Su R, *et al*. Prediction of human protein subcellular localization using deep learning. *J Parallel Distrib Comput* 2018;**117**:212.
51. Wei L, Su R, Wang B, *et al*. Integration of deep feature representations and handcrafted features to improve the prediction of n 6-methyladenosine sites. *Neurocomputing* 2019;**324**:3.
52. Su R, Wu H, Xu B, *et al*. Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**16**:1231.
53. Dao FY, Lv H, Yang YH, *et al*. Computational identification of n6-methyladenosine sites in multiple tissues of mammals. *Comput Struct Biotechnol J* 2020;**18**:1084.
54. Shen Y, Tang J, Guo F. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into chou's general pseaac. *J Theor Biol* 2019;**462**:230.
55. Shen Y, Ding Y, Tang J, *et al*. Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief Bioinform* 2019. doi: 10.1093/bib/bbz106.
56. Cabarle FGC, de la Cruz RTA, Zhang X, *et al*. On string languages generated by spiking neural p systems with structural plasticity. *IEEE Trans Nanobiosci* 2018;**17**:560.
57. Bateman A, Martin M-J, Orchard S, *et al*. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506.
58. El-Gebali S, Mistry J, Bateman A, *et al*. The pfam protein families database in 2019. *Nucleic Acids Res* 2019;**47**:D427.
59. Bengio Y. Deep learning of representations for unsupervised and transfer learning. In: Isabelle G, Gideon D, Vincent L *et al*. (eds). *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. Proceedings of Machine Learning Research: PMLR, 2012, 17.
60. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. 2018; arXiv:1802.03426.
61. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances In Neural Information Processing Systems*, Vol. **30**. La Jolla: Neural Information Processing Systems (Nips), 2017.
62. Chen Z, Zhao P, Li F, *et al*. Ilearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, rna and protein sequence data. *Brief Bioinform* 2020;**21**:1047.
63. Liu B. Bioseq-analysis: a platform for DNA, rna and protein sequence analysis based on machine learning approaches. *Brief Bioinform* 2019;**20**:1280.
64. Liu B, Liu F, Wang X, *et al*. Pse-in-one: a web server for generating various modes of pseudo components of DNA, rna, and protein sequences. *Nucleic Acids Res* 2015;**43**:W65.
65. Chen Z, Zhao P, Li F, *et al*. Ifeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499.
66. Tang Y-J, Pang Y-H, Liu B. Idp-seq2seq: identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformaitcs* . doi: 10.1093/bioinformatics/btaa667.
67. Basith S, Manavalan B, Shin TH, *et al*. Sdm6a: a web-based integrative machine-learning framework for predicting 6ma sites in the rice genome. *Mol Ther Nucl Acids* 2019;**18**:131.
68. Manavalan B, Basith S, Shin TH, *et al*. Meta-4mcpred: a sequence-based meta-predictor for accurate DNA 4mc site

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

prediction using effective feature representation. *Mol Ther Nucl Acids* 2019;**16**:733.
69. Dhall A, Patiyal S, Sharma N, *et al*. Computer-aided prediction and design of il-6 inducing peptides: Il-6 plays a crucial role in covid-19. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa259.
70. Dwivedi VD, Arya A, Yadav P, *et al*. Denvind: dengue virus inhibitors database for clinical and molecular research. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa098.
71. Ding H, Yang W, Tang H, *et al*. Phypred: a tool for identifying bacteriophage enzymes and hydrolases. *Virol Sin* 2016;**31**:350.
72. Tang H, Zhao YW, Zou P, *et al*. Hbpred: a tool to identify growth hormone-binding proteins. *Int J Biol Sci* 2018;**14**:957.
73. Lv Z, Wang D, Ding H, *et al*. Escherichia coli DNA n-4-methycytosine site prediction accuracy improved by light gradient boosting machine feature selection technology. *IEEE Access* 2020;**8**:14851.
74. Lv Z, Ding H, Wang L, *et al*. A convolutional neural network using dinucleotide one-hot encoder for identifying DNA n6-methyladenine sites in the rice genome. *Neurocomputing* 2021;**422**:214.
75. Zhang SC, Li XL, Zong M, *et al*. Efficient knn classification with different numbers of nearest neighbors. *IEEE Trans Neural Netw Learn Syst* 2018;**29**:1774.
76. Du L, Meng QF, Chen YH, *et al*. Subcellular location prediction of apoptosis proteins using two novel feature extraction methods based on evolutionary information and lda. *Bmc Bioinf* 2020;**21**:19.
77. Capellini TD, Vaccari G, Ferretti E, *et al*. Scapula development is governed by genetic interactions of pbx1 with its family members and with emx2 via their cooperative control of alx1. *Development* 2010;**137**:2559.
78. Zhu XJ, Feng CQ, Lai HY, *et al*. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl-Based Syst* 2019;**163**:787.
79. Tan JX, Li SH, Zhang ZM, *et al*. Identification of hormone binding proteins based on machine learning methods. *Math Biosci Eng* 2019;**16**:2466.
80. Huo Y, Xin L, Kang C, *et al*. Sgl-svm: a novel method for tumor classification via support vector machine with sparse group lasso. *J Theor Biol* 2020;**486**:110098.
81. Wang Y, Liu K, Ma Q, *et al*. Pancreatic cancer biomarker detection by two support vector strategies for recursive feature elimination. *Biomark Med* 2019;**13**:105.
82. Lv Z, Jin S, Ding H, *et al*. A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. *Front Bioeng Biotechnol* 2019;**7**:215.
83. Liu B, Li K. Ipromoter-2l2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol Ther-Nucl Acids* 2019;**18**:80.
84. Lv H, Dao FY, Zhang D, *et al*. Idna-ms: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* 2020;**23**:100991.
85. Wang M, Yue L, Cui X, *et al*. Prediction of extracellular matrix proteins by fusing multiple feature information, elastic net, and random forest algorithm. *Mathematics* 2020;**8**:169.
86. Wang X, Yu B, Ma A, *et al*. Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* 2019;**35**:2395.
87. Shi H, Liu S, Chen J, *et al*. Predicting drug-target interactions using lasso with random forest based on evolutionary information and chemical structure. *Genomics* 2019;**111**:1839.
88. Zhang YJ, Yu S, Xie RP, *et al*. Pengaroo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics* 2020;**36**:704.
89. Yu JW, Xuan ZW, Feng X, *et al*. A novel collaborative filtering model for lncrna-disease association prediction based on the naive bayesian classifier. *Bmc Bioinf* 2019;**20**:13.
90. Chen K, Wei Z, Zhang Q, *et al*. Whistle: a high-accuracy map of the human n-6-methyladenosine (m(6)a) epitranscriptome predicted using a machine learning approach. *Nucl Acids Res* 2019;**47**:e41. https://doi.org/10.1093/nar/gkz074.
91. Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant Biol* 2016;**4**:320.
92. Ma AJ, Wang CK, Chang YZ, *et al*. Iris3: integrated cell-type-specific regulon inference server from single-cell rna-seq. *Nucl Acids Res* 2020;**48**:W275.
93. Wei L, Wan S, Guo J, *et al*. A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif Intell Med* 2017;**83**:82.
94. Wei L, Xing P, Zeng J, *et al*. Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif Intell Med* 2017;**83**:67.
95. Shen C, Jiang L, Ding Y, *et al*. Lpi-ktaslp: prediction of lncrna-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* 2019;**7**:13486.
96. Ding Y, Tang J, Guo F. Identification of drug-side effect association via semi-supervised model and multiple kernel learning. *IEEE J Biomed Health Inform* 2019;**325**:211.
97. Ding Y, Tang J, Guo F. Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 2019;**325**:211.
98. Zeng X, Zhu S, Liu X, *et al*. Deepdr: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019;**35**:5191.
99. Cheng L, Zhuang H, Ju H, *et al*. Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front Genet* 2019;**10**:94.
100. Cheng L, Qi C, Zhuang H, *et al*. Gutmdisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucl Acids Res* 2020;**48**:D554.
101. Chan KKK, Zhang J, Chia NY, *et al*. Klf4 and pbx1 directly regulate nanog expression in human embryonic stem cells. *Stem Cells* 2009;**27**:2114.
102. Basith S, Manavalan B, Hwan Shin T, *et al*. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med Res Rev* 2020;**40**:1276.
103. Hasan MM, Basith S, Khatun MS, *et al*. Meta-i6ma: an interspecies predictor for identifying DNA n6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa202.
104. Hasan MM, Manavalan B, Shoombuatong W, *et al*. I6ma-fuse: improved and robust prediction of DNA 6 ma sites in the rosaceae genome by fusing multiple feature representation. *Plant Mol Biol* 2020. doi: 10.1007/s11103-020-00988-y:10.1007/s11103.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021

105. Li XY, Li WK, Zeng M, *et al*. Network-based methods for predicting essential genes or proteins: a survey. *Brief Bioinform* 2020;**21**:566.
106. Wang DL, Liu DP, Yuchi JK, *et al*. Musitedeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res* 2020;**48**:W140.
107. Armenteros JJA, Tsirigos KD, Sonderby CK, *et al*. Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019;**37**:420.
108. Lu P, Min D, DiMaio F, *et al*. Accurate computational design of multipass transmembrane proteins. *Science* 2018;**359**:1042.
109. Fu X, Cai L, Zeng X, *et al*. Stackcppred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics* 2020;**36**:3028.
110. Song B, Li K, Orellana-Martín D, *et al*. Cell-like p systems with evolutional symport/antiport rules and membrane creation. *Inf Comput* 2020;**104542**.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab008/6126754 by University of Cambridge user on 28 May 2021