ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Active Semisupervised Model for Improving the Identification of Anticancer Peptides

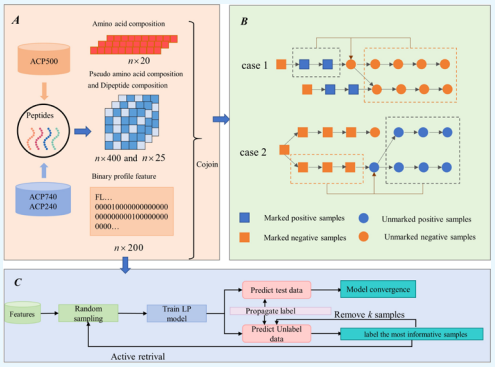
Lijun Cai, Li Wang, Xiangzheng Fu,* and Xiangxiang Zeng

Cite This: *ACS Omega* 2021, 6, 23998−24008

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Cancer is one of the most dangerous threats to human health. Accurate identification of anticancer peptides (ACPs) is valuable for the development and design of new anticancer agents. However, most machine-learning algorithms have limited ability to identify ACPs, and their accuracy is sensitive to the amount of label data. In this paper, we construct a new technology that combines active learning (AL) and label propagation (LP) algorithm to solve this problem, called (ACP-ALPM). First, we develop an efficient feature representation method based on various descriptor information and coding information of the peptide sequence. Then, an AL strategy is used to filter out the most informative data for model training, and a more powerful LP classifier is cast through continuous iterations. Finally, we evaluate the performance of ACP-ALPM and compare it with that of some of the state-of-the-art and classic methods; experimental results show that our method is significantly superior to them. In addition, through the experimental comparison of random selection and AL on three public data sets, it is proved that the AL strategy is more effective. Notably, a visualization experiment further verified that AL can utilize unlabeled data to improve the performance of the model. We hope that our method can be extended to other types of peptides and provide more inspiration for other similar work.



1. INTRODUCTION

In recent years, the incidence and mortality of cancer have shown a gradual upward trend, and it is one of the main diseases threatening human life. Although cancer can be treated using traditional physical and chemical methods, including targeted therapy, chemotherapy, and radiation therapy, these methods are expensive and inefficient.[1,2] Besides, some anticancer drugs have shown adverse effects on normal cells, and cancer cells can develop drug resistance.[3−6] Therefore, the discovery and rational design of more effective therapeutic drugs are urgently needed. Anticancer peptides (ACPs) are usually short peptides with a length of 5−30 amino acids and are natural agents with high efficacy, selectivity, and specificity; as such, they have been widely recognized as one of the safest and most reliable anticancer therapeutics over the years.[7] Currently, a large number of ACP-based drugs are being evaluated in various stages of clinical trials.[8,9] In this context, identifying ACPs from large-scale protein sequences is crucial.[10,11]

Unfortunately, wet-laboratory experimental identification and development of novel ACPs are extremely cost-ineffective and time-consuming.[12,13] Therefore, more and more researchers focus on developing data-driven computational methods, such as machine learning (ML), to identify ACPs. According to the key technology of ML, mainstream computational identification methods of ACPs can be usually divided into two categories: mining feature information and designing efficient classifiers. The earliest, developed by Tyagi et al.,[14] is an ACP predictor that uses amino acid and dipeptide composition (AAC and DC). Subsequently, Hajisharifi et al.[15] proposed a new feature representation method, which not only includes pseudo-amino acid composition but also increases the local correlation and sequence information of residue. Recently, Wei et al.[4] extracted feature descriptors about the amino acid composition and physical and chemical properties of ACPs and achieved satisfactory performance. In another work, Rao et al.[7] proposed the use of multiview information to further improve the feature representation of the learning scheme, which is significantly superior to the existing prediction tools.

In addition to the above-mentioned methods of mining ACP feature information, another part of the research involves exploring potential classifiers. For example, Manavalan et al.[16] applied support vector machine (SVM) and random forest (RF) to identify ACPs. You et al.[17] implemented a deep long

Received: June 15, 2021
Published: September 8, 2021

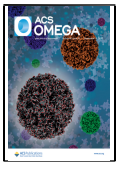


Table 1. Influence of Parameter Gamma on the Model Performance

gamma	0.20	0.22	0.24	0.26	0.28	0.30	0.32	0.34	0.36	0.38
ACC	0.80	0.84	0.78	0.86	0.82	0.92	0.85	0.83	0.84	0.88
AUC	0.76	0.75	0.73	0.81	0.86	0.88	0.76	0.77	0.83	0.85

short-term memory (LSTM) model to identify ACPs and non-ACPs. In addition, Liang et al.[18] improved potential models by conducting comparative experiments on classic models, such as SVM, Naive Bayesian, Light Gradient Boosting Machine (lightGBM), and k-nearest neighbors (KNNs). Recently, Muhammod et al.[19] proposed a new multihead deep convolutional neural network model, which further leveraged deep learning for ACP identification and obtained excellent experimental results. The last few years have witnessed the development of computation-based methods, especially related to ML.

Although ML has achieved considerable success, the identification of ACPs still presents non-negligible challenges. First, ACPs are too short to capture specific information, and straightforward integration of various types of feature descriptors leads to dimensional disasters. Second, most established ML models usually face the dilemma of data starvation and require a large amount of expensive labeled data. Finally, further improving the accuracy and robustness of ACP identification is necessary to realize their real medical applications. Fortunately, through selecting the data points whose labels would be most informative in the learning task, AL not only addresses data deficiency but also improves the model accuracy, and it has been successfully applied in many tasks.[20−28] Combinations of AL and semisupervised techniques have been proposed in the literature and are somewhat prosperous when applied to a particular context.[29−33] In view of the promising potential shown by AL and semisupervised scheme, they are expected to solve the issues of ACP identification.

Therefore, we designed a novel framework called the **a**ctive **l**abel **p**ropagation **m**odel (ACP-ALPM) for ACP identification by taking advantage of both labeled and unlabeled peptides. First, we designed a novel feature representation method of ACPs, which not only contains sequence order, local correlation, and residue information but also supplements efficient one-hot encoding information. Thereafter, we introduced a semisupervised label propagation (LP) algorithm as a benchmark model and incorporated AL strategies to continuously update and optimize the model by iteratively choosing the most informative data. Finally, we compared ACP-ALPM with 12 advanced methods and 6 well-designed methods, and the results showed that our method had advantages in identifying ACPs. In addition, the implementation of ablation analysis on three data sets showed that AL played an important role in ACP-ALPM.

2. RESULTS

2.1. Evaluation Metrics. To measure the performance, we selected seven classic metrics widely used in two-class identification problems, including accuracy (Acc), recall, sensitivity (Sens), specificity (Spec), precision (Prec), F1-score, and Matthews correlation coefficient (MCC). They are calculated as follows

$$\begin{cases} \text{Acc} = \dfrac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \\ \text{recall} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Sens} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Spec} = \dfrac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{Prec} = \dfrac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{F1} - \text{score} = \dfrac{2 \times \text{recall} \times \text{Prec}}{\text{recall} + \text{Prec}} \\ \text{MCC} = \dfrac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \end{cases} \quad (1)$$

where TN is the true negative number, TP is the true positive number, FN is the false negative number, and FP is the false positive number. The area under the curve (AUC) was also adopted to evaluate the performance of the model.

2.2. Experimental Setup. Our model had a key parameter α, which controlled the radial distance range. Thus far, the parameter in our experiments was tuned based on the performance over a hold-out set. We tested the sensitivity of the model to the choice of parameter. As shown in Table 1, when the parameter α was 0.3, the model had the best effect. In addition, we heuristically set the parameter N to 50 and k to 5. Then, we discussed the influence of parameter epoch on the model. Figure 1 shows that when the epoch is too small, the accuracy rate is not optimal, and when the epoch is too large, the accuracy rate begins to decrease. When the epoch is 13, the

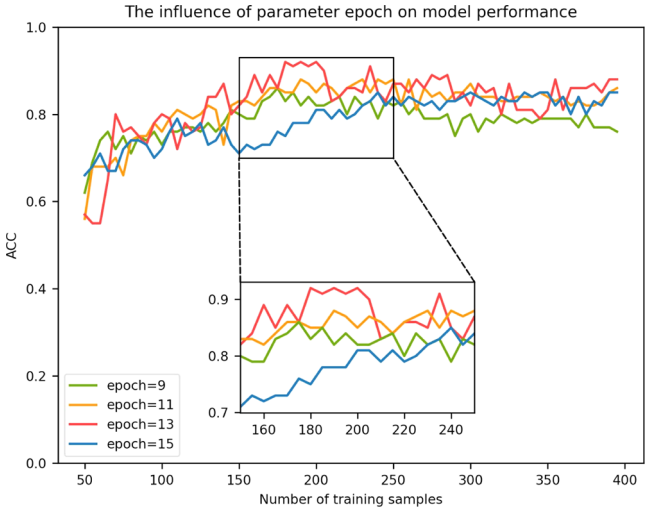


Figure 1. Influence of parameters epoch on ACC of ACP-ALPM.

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

training data interval is 180−200, and the accuracy rate is optimal.

In addition, we randomly split 80% of our data into the training set and 20% into the test set. To prove the generalization of the model, we also conducted comparative experiments on additional data sets ACP740 and ACP240.

2.3. Performance Comparison with Different Classifiers. To evaluate the identification ability of the proposed ACP-ALPM, we compared the performances of six different classifiers on the benchmark data set ACP500, three of which were classic deep-learning and ML models, and the three others combined the AL strategy based on selected comparison models. In detail, we adopt the ACP-ALPM scheme but replace the label propagation network with adaptive boosting (Adaboost), deep neural networks (DNNs), and k-nearest neighbors (KNNs); finally, we obtained AAdaboost, ADNN, and AKNN as complementary comparison models. The experimental comparison results of ACP-ALPM and the six other methods are reported in Table 2 and Figure 2A.

Table 2. Comparison of ACP-ALPM and the Other Six Methods on ACP500

methods	Acc	recall	Prec	F1-score	AUC
DNN	0.850	0.880	0.830	0.854	0.850
ADNN	0.870	0.851	0.870	0.860	0.869
KNN	0.830	0.855	0.869	0.862	0.822
AKNN	0.880	0.881	0.912	0.897	0.880
Adaboost	0.830	0.870	0.825	0.847	0.826
AAdaboost	0.870	0.850	0.927	0.887	0.875
ACP-ALPM	0.920	0.981	0.883	0.930	0.915

As shown in Figure 2A and Table 2, the performance of the proposed ACP-ALPM was significantly better than that of the other models. Specifically, the values of Acc, recall, Prec, F1-score, and AUC of our method were 0.920, 0.957, 0.882, 0.918, and 0.922, respectively. Compared with models without AL, the values of Acc, recall, Prec, F1-score, and AUC increased by 7.0−9.0, 7.7−9.8, 5.2−5.7, 6.4−7.1, and 7.2%−10.0%, respectively. Our model was superior to the models with AL; the values of Acc and AUC improved by 3.0−4.0 and 4.2−5.3%, respectively. The model that uses AL is generally superior to the ordinary model; for instance, the values of Acc, Prec, F1-score, and AUC of ADNN are 0.870, 0.870, 0.860, and 0.869, respectively, which are 2.0, 4.0, 0.6, and 1.9% higher than those of DNN. The results indicate that ACP-ALPM has a stronger ability than the other classifiers for identifying true ACPs from non-ACPs, and AL provides a vital contribution to identification performance.

2.4. Performance Comparison with State-of-the-Art Methods. To validate the superiority of ACP-ALPM, we compared its performance with some state-of-the-art methods, including AntiCP,[14] Hajisharifi's method,[15] iACP,[34] ACPred-FL,[4] DeepACP,[35] PEPred-Suite,[36] ACPred-Fuse,[7] ACP-DL,[17] and ACP-MHCNN.[19] Among them, AntiCP represents two predictors composed of amino acids and dipeptides, and DeepACP represents the classifiers generated by the recurrent neural network. For a fair comparison, all approaches were trained and tested on the ACP500 data sets. Table 3 illustrates the predictive performance in terms of five metrics (Sens, Spec, Acc, MCC, and AUC) on this data set, and Figure 3A shows the overall effect of considering these five performances. As shown in Table 3 and Figure 3A, the performance of the proposed ACP-ALPM was significantly better than those of the

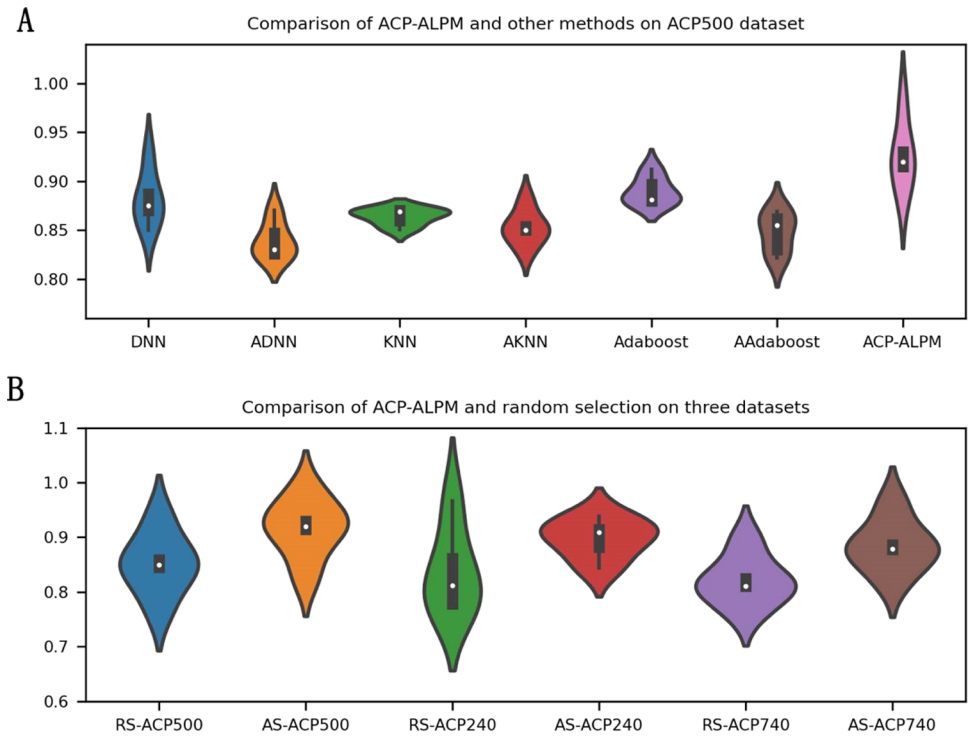


Figure 2. Multiangle comparative experiment of ACP-ALPM. (A) Results of the ACP500 data set for overall performance comparison of ACP-ALPM and different classifiers (DNN, KNN, Adaboost, DNN with active strategy, KNN with active strategy, Adaboost with active strategy). (B) Overall performance comparison of the proposed active selection (AS) strategy of ACP-ALPM and random selection (RS) scheme on ACP500, ACP240, and ACP740 data sets.

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

Table 3. Performance Comparisons of our Proposed ACP-ALPM with the Existing Methods

methods	Sens	Spec	Acc	MCC	AUC
AntiCP_AAC	0.668	0.784	0.726	0.455	0.824
AntiCP_DC	0.716	0.776	0.746	0.493	0.825
Hajisharifi et al.	0.672	0.836	0.754	0.515	0.831
iACP	0.572	0.840	0.706	0.428	0.809
ACPred-FL	0.716	0.844	0.780	0.565	0.846
CNN-RNN	0.720	0.817	0.768	0.539	0.871
CNN	0.780	0.793	0.786	0.573	0.903
DeepACP(RNN)	0.780	0.878	0.829	0.662	0.920
PEPred-Suite	0.728	0.880	0.804	0.615	0.860
ACPred-Fuse	0.772	0.876	0.824	0.652	0.882
ACP-DL	0.890	0.805	0.847	0.620	
ACP-MHCNN	0.976	0.842	0.910	0.820	
ACP-ALPM	0.981	0.848	0.920	0.940	0.915

other methods. For example, the values of Sens, Acc, and MCC of our ACP-ALPM were 0.981, 0.920, and 0.940, respectively, which were 5.0−40.9, 1.0−21.4, and 12.0−51.2%, respectively. Although our AUC was not the highest, it was only slightly lower (by 0.5%) than the best AUC of DeepACP. Additionally, our method's Spec indicator was better than the majority of the methods. As such, ACP-ALPM can identify ACPs more precisely than the existing methods. In addition, our method has an advantage in running speed, and the experiment takes about 13 s.

2.5. Contribution Analysis for Different Sequence Representations. To demonstrate that our method has abundant and effective feature information to achieve better performance, we conducted feature contribution analysis. Under this experimental setting, we compared our method with four different combination experiments, and the corresponding results are reported in Table 4 and Figure 3B.

Table 4. Five Combinations of Four Sequence Representations Explored in This Research[a]

	combination	Acc	recall	Prec	F1-score	AUC
1	AAC + DC + PC-PseAAC	0.720	0.855	0.701	0.770	0.705
2	AAC + DC + BPF	0.890	0.878	0.896	0.887	0.890
3	AAC + PC-PseAAC + BPF	0.900	0.983	0.864	0.919	0.884
4	DC + PC-PseAAC + BPF	0.910	0.957	0.863	0.907	0.913
5	AAC + DC + PC-PseAAC + BPF	0.920	0.981	0.883	0.930	0.915

[a]*Note*: no. 5 represents the feature representation of our proposed method.

As shown in Table 4, for the ACP500 data set, among the four representation combinations (nos. 1, 2, 3, 4), the representation binary profile feature (BPF) information contributes more to our model than AAC, DC, and parallel correlation pseudo-amino acid composition (PC-PseAAC) information, which shows the ability of BPF to capture short peptide sequences and distinguish ACPs. Figure 3B shows that the feature representation of our proposed method (no. 5) performs better than those of the four other feature representation methods on overall evaluation metrics.

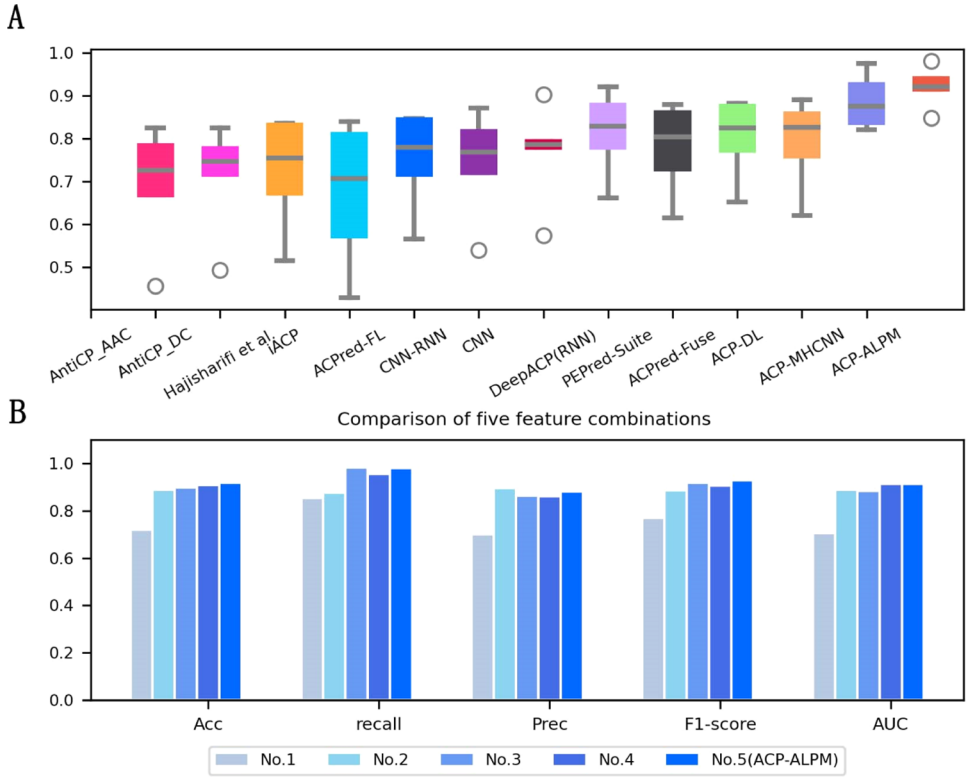


Figure 3. Comparative experiment of ACP-ALPM on ACP500. (A) Overall performance comparison of the proposed ACP-ALPM and 12 state-of-the-art predictors on the ACP500 data set. (B) Overall performance comparison of five different feature combination methods, among which no. 5 is the method adopted by ACP-ALPM.

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

2.6. Comparison Experiments with Random Data Selection. The AL strategy of ACP-ALPM requires labels iteratively based on the distribution of peptides instances and the learned decision function that is refined at each iteration. The random selection strategy simply requests labels through a method of random shuffling and reslicing. To illustrate the effectiveness and robustness of our strategy, we conducted a comparative experiment of active selection (AS) and random selection (RS) on three benchmark data sets (ACP500, ACP240, and ACP740). The experimental results are presented in Table 5 and Figure 2B.

Table 5. Comparison of ACP-ALPM and Random Selection on the ACP500 Data Set[a]

	Acc	recall	Prec	F1-score	AUC
RS-ACP240	0.812	0.966	0.778	0.862	0.772
AS-ACP240	0.915	0.938	0.882	0.909	0.844
RS-ACP740	0.811	0.892	0.767	0.825	0.811
AS-ACP740	0.878	0.959	0.824	0.886	0.879
RS-ACP500	0.850	0.932	0.774	0.845	0.859
AS-ACP500	0.920	0.981	0.833	0.930	0.915

[a]*Note*: AS-ACP500, AS-ACP240, and AS-ACP240 use the AL strategy of ACP-ALPM.

It can be seen from Table 5 and Figure 2B that in the three data sets, the performance of the AS we adopted is significantly better than that of RS as a whole. Although the value of recall of AS-ACP240 was not as high as that of RS-ACP240, it was slightly worse than that of RS-ACP240. It should be noted that RS-ACP240 means the method of adopting a random selection (RS) strategy on the ACP240 data set. Therefore, it can be concluded that the data selection strategy is closely associated with the model identification performance. Active learning strategies can filter out information-rich data for the model, and a stronger classifier can be built; random selection strategies may generate more noisy data, which will affect the quality of model identification.

2.7. Impact of AL on ACP-ALPM. Additional evaluation comparison experiments were performed on ACP500, ACP240, and ACP740 to demonstrate the necessity of AL and the universal superiority of ACP-ALPM to other data sets. The comparison between ACP-ALPM and non-ACP-ALPM is presented in Table 6. The key difference between non-ACP-ALPM and ACP-ALPM is the absence of an AL strategy, that is, the training set data is fed all at once.

Table 6 shows the experimental results of the ACP500, ACP240, and ACP740 data sets. The five evaluation indicators (Acc, recall, Prec, F1-score, and AUC) indicate that ACP-ALPM outperforms the non-ACP-ALPM method on the benchmark data set ACP500. In addition to the superior performance on the benchmark data set, the experimental results of the additional two data sets again reflect the contribution of AL to the model. Compared with the evaluation values of non-ACP-ALPM, the value of Acc of ACP-ALPM was increased by 6.3 and 5.3% on the ACP240 and ACP740, respectively; the value of recall of ACP-ALPM was increased by 0.7 and 7.8% on the ACP240 and ACP740, respectively; the value of Prec of ACP-ALPM was increased by 8.8 and 1.7% on the ACP240 and ACP740, respectively; the F1-score of ACP-ALPM was increased by 5.2 and 4.1% on the ACP240 and ACP740, respectively; and the value of AUC of ACP-ALPM was increased by 6.3 and 6.0% on the ACP240 and ACP740, respectively. These outstanding results exhibit the generalization ability and robustness of the AL strategy of ACP-ALPM.

2.8. Visualization Analysis of ACP-ALPM. Our method is based on uncertain sampling iterations of selecting data nodes from the data pool and learning classification through the propagation of labels between similar nodes. To explain the active label propagation mechanism between data nodes, we visualized the distribution and variation process of nodes, as shown in Figure 4. Blue represents positive samples, magenta represents negative samples, squares represent marked data, and circles represent unmarked data. Labeled data can propagate labels to unlabeled data, which increases the number of labeled data.

As shown in Figure 4, first, it can be observed that the overall data points are increasing, which indicates that active learning is filtering more data for the label propagation network to learn. Secondly, by observing each subgraph in order, we can find that the square data points obviously increase gradually, and the circular data points decrease, which means that the unlabeled samples are marked with the corresponding labels. Finally, observing Figure 4A−D, we found that there are obvious changes between the pictures, and the experimental performance is indeed significantly improved. Observing Figure 4D−F, we find that the picture changes are no longer obvious, and the model tends to converge.

In addition, the regional distribution of positive and negative samples has obvious distances, which verifies that AL can utilize unlabeled data to assist the LP network in effective labeling.

3. DISCUSSION

Currently, many peptide-based therapies are being evaluated in terms of their efficacy to treat various tumor types across different phases of preclinical and clinical trials, resulting in peptides becoming an important alternative anticancer therapeutic agent. Here, we proposed a calculation method for the identification of anticancer peptides based on AL, called ACP-ALPM. Unlike traditional machine-learning methods, AL can make full use of unlabeled data and wisely choose the most informative data. Coupled with the label propagation algorithm to generate a powerful anticancer peptide classifier.

We selected some state-of-the-art methods for experimental comparison, which only extracted the information of labeled data, and the experimental results showed that our method had

Table 6. Comparison of ACP-ALPM and Non-ACP-ALPM on Three Benchmark Data Sets

	ACP-ALPM					non-ACP-ALPM				
data sets	Acc	recall	Prec	F1-score	AUC	Acc	recall	Prec	F1-score	AUC
ACP500	0.920	0.981	0.883	0.930	0.915	0.870	0.880	0.863	0.871	0.870
ACP240	0.875	0.938	0.882	0.909	0.844	0.812	0.931	0.794	0.857	0.781
ACP740	0.878	0.959	0.824	0.886	0.879	0.825	0.887	0.807	0.845	0.819

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

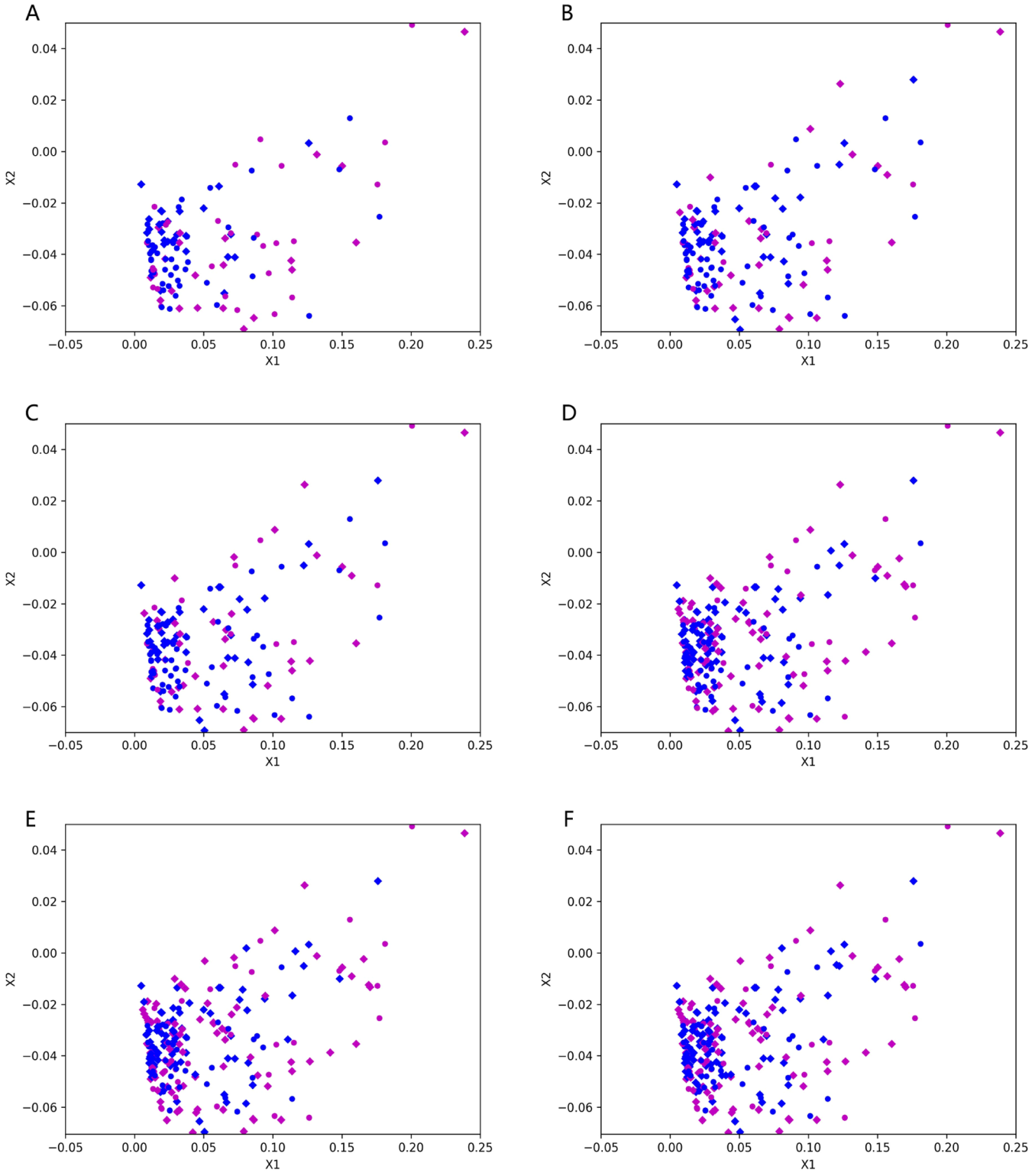


Figure 4. Visualization results (A–F) on the benchmark data set ACP500 of active label propagation.

better performance in identifying anticancer peptides. More importantly, extensive benchmark tests show that compared with randomly selected label propagation algorithms, AL strategies contribute more to ACP-ALPM, which strongly demonstrates the effectiveness of fusion AL strategies. In addition, we visualized the process of ACP-ALPM data selection and label propagation and increased the interpretability of the model's operating mechanism.

In short, ACP-ALPM is a powerful method that can be extended to a wide range of biochemical applications with a simple adjustment of parameters and can be used to identify the types of training data. In the future, we envision using more efficient algorithms to optimize the data screening process of active learning, such as clustering algorithms to further improve the performance of the model.

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

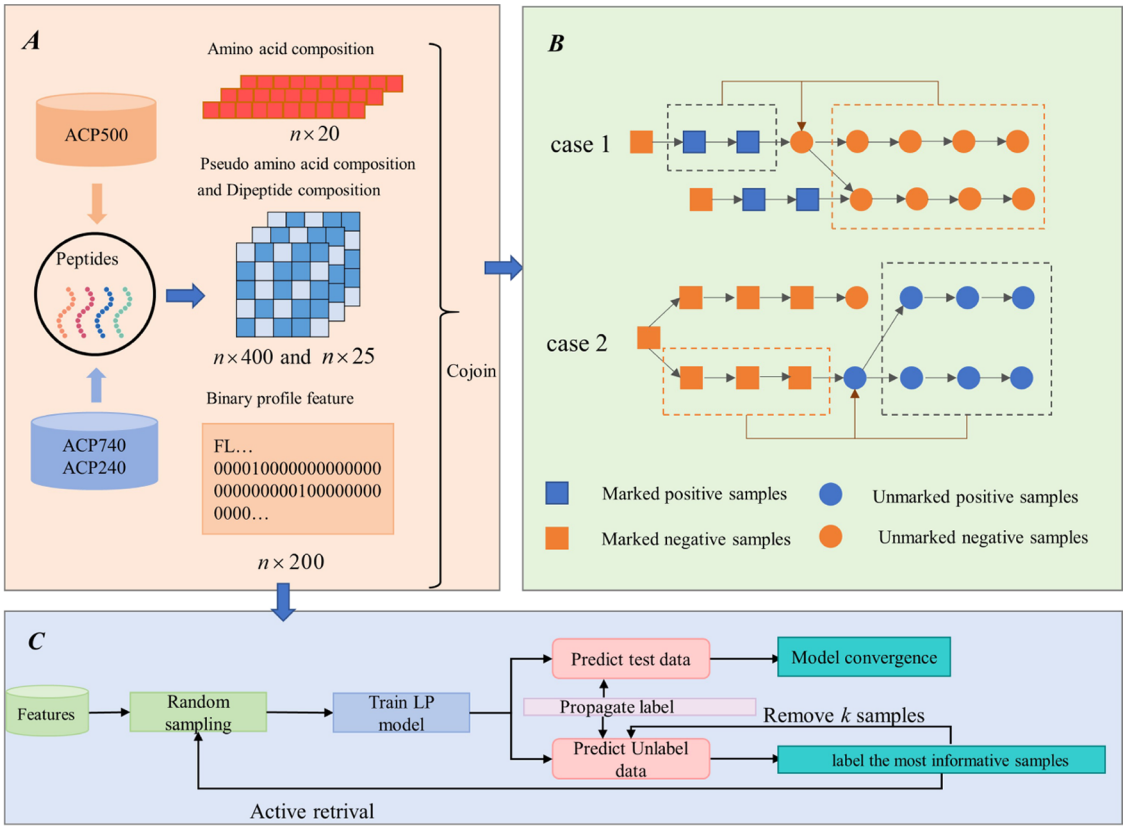


Figure 5. Overall framework of ACP-ALPM. (A) Construct distinctive feature vectors on three public data sets. (B) Label of the current unlabeled node (positive sample or negative sample) is determined by the labeled sample (positive sample and negative sample) and the unlabeled sample (positive sample and negative sample). (C) Label propagation process integrated with AL. First, we train an initial LP model based on uncertain sampling and predict the data to gain the labeling results. Then, according to the probability distributions provided by the classifier, we select the most informative data to feedback to the training set and repeat the entire process model until the model converged.

4. CONCLUSIONS

In this article, we have proposed a new framework that combines AL strategies and LP algorithms to identify anticancer peptides more accurately and efficiently than the latest methods.

Specifically, we propose a novel feature representation scheme to effectively represent ACP sequences, which includes sequence-based feature descriptors (AAC and DC), local correlation of residues' information (PC-PseAAC), and abundant one-hot encoding information (BPF). Then, we design the active label propagation network to learn feature information, which is a graph-based semisupervised learning method to obtain the label information of unlabeled nodes from the label information of the labeled nodes, and integrate AL strategies to reduce the probability of error propagation. Experimental results demonstrated that (1) in terms of the presented evaluation indicators, ACP-ALPM is superior to the state-of-the-art and classic methods in identifying anticancer peptides; (2) our proposed feature combination representation leads to better identification performance; and (3) compared with random selection and not using AL strategies, LP algorithms integrated with AL have stronger identification capabilities. Overall, ACP-ALPM is an efficient, robust, and generalizable method.

As an implementation of this work, we have also made a free and open code of ACP-ALPM for the wider research community to use. ACP-ALPM can deal with the identification of other peptides, and only a simple parameter adjustment can achieve good results. In the future, we will try to explore multidimensional or multiangle anticancer peptide characteristics and optimize active learning strategies to better integrate with advanced and popular frameworks.

5. MATERIALS AND METHODS

In this section, we present a description of the framework of ACP-ALPM, and the specific workflow is shown in Figure 5.

5.1. Data Set. In this study, we used two groups of ACP data sets from the existing literature to evaluate the ACP-ALPM performance. A set of data sets designed to train the target model and test the performance of the model was named ACP500, and other data sets used to prove the universal advantages of ACP-ALPM were named ACP740 and ACP240. The amino acid distributions of the three benchmark data sets are shown in Figure 6.

For the first data set ACP500, we randomly sampled the benchmark data set established by Wei et al.[4] and obtained 250 experimentally validated positive samples and 250 negative samples. The positive and negative samples of these data are balanced and underwent CD-HIT[37] deredundancy processing.

We further constructed the second data set downloaded from the study of Yi et al.[17] ACP740 and ACP240 contained 376 and 129 positive samples and 364 and 111 negative samples, respectively. Similarly, these data sets used the tool CD-HIT[37] to remove sequences with more than 90% similarity.

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

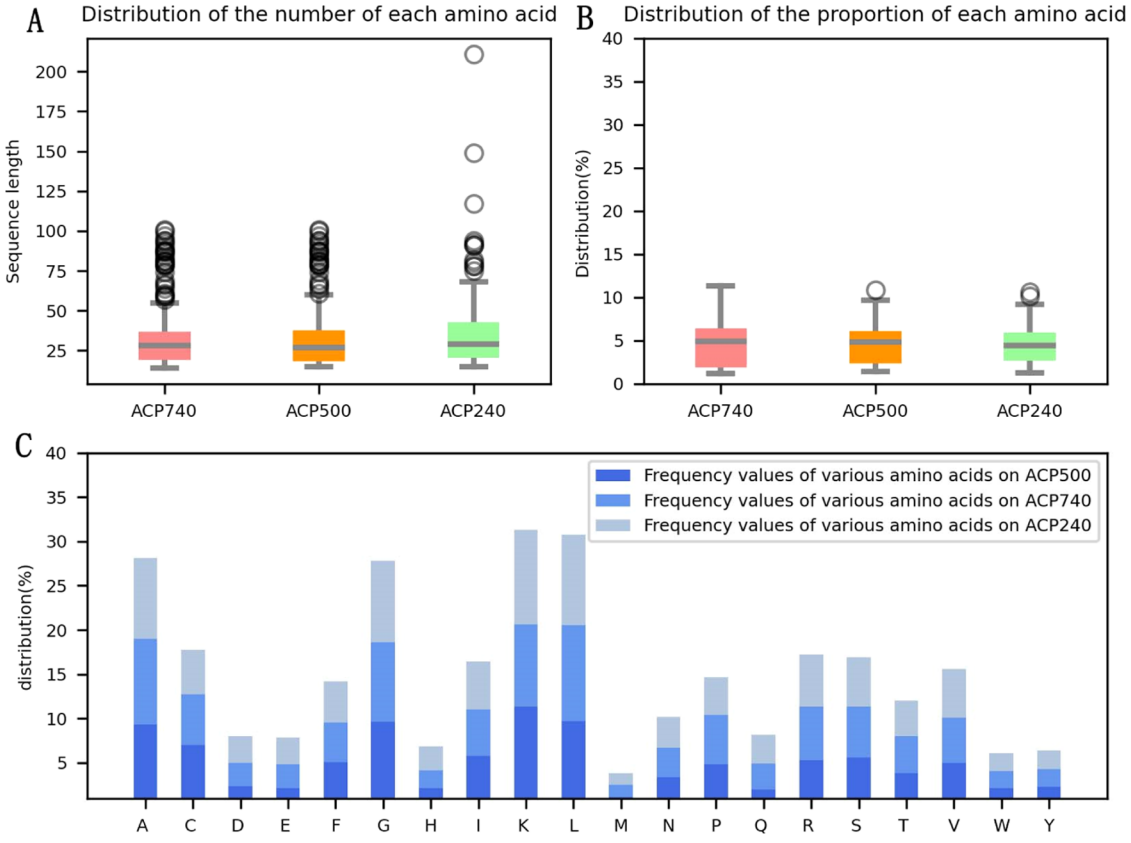


Figure 6. Amino acid distribution map of ACP500, ACP740, and ACP240. (A) Distribution of the number of 20 amino acids on three different data sets. (B) Distribution of the proportion of 20 amino acids on three different data sets. (C) Statistics of amino acid distribution on three different data sets.

5.2. Feature for Peptide Sequence. *5.2.1. Amino Acid Composition (AAC).* AAC[38] is a means to calculate the frequency of each type of amino acid in a protein or peptide sequence, which is represented by characters as shown in the formula

$$F_1(P) = (f_1, f_2, f_3, \cdots f_{20}) \tag{2}$$

$$f_i = \frac{c_i}{l} \tag{3}$$

where c_i is the number of type i appearing in the peptide, l is the length of the peptide, and f_i is the percent composition of amino acid type i. The dimension of the AAC descriptor is 20.

5.2.2. Dipeptide Composition (DC). DC[38] refers to the frequency at which two amino acids in a protein sequence constitute a dipeptide. It is defined as

$$F_2(\alpha, \beta) = \frac{N_{\alpha\beta}}{N-1}, (\alpha, \beta \in \{A, C, D, \cdots, Y\}) \tag{4}$$

where $N_{\alpha\beta}$ is the number of dipeptides represented by amino acid types α and β and N is the total number of amino acids in the character sequence of a given peptide. The dimension of the DC descriptor is 400.

5.2.3. Parallel Correlation Pseudo-Amino Acid Composition (PC-PseAAC). PC-PseAAC is a pseudo-amino feature extraction method based on protein molecular character sequence first proposed by Chou.[39] Given any protein sequence P, it can be numerically represented by a feature vector composed of 20+ λ-dimensional pseudo-amino acids as indicated by formula 5

$$P = [r_1, r_2, r_3, \cdots, r_{20}, r_{20+1}, \cdots, r_{20+\lambda}] \tag{5}$$

where

$$r_k = \begin{cases} \dfrac{f_k}{\sum_{\alpha=1}^{20} f_\alpha + w\sum_{\beta=1}^{\lambda}\delta_\beta} & 1 \le k \le 20 \\ \dfrac{w\delta_{k-20}}{\sum_{\alpha=1}^{20} f_\alpha + w\sum_{\beta=1}^{\lambda}\delta_\beta} & 21 \le k \le 20+\lambda \end{cases} \tag{6}$$

where f_α is the occurrence frequency of the 20 native amino acids in the peptide; w is the weight factor ranging from 0 to 1; and δ_β is the correlation factor of the β-tier residue of given P protein sequence that is defined as

$$\delta_\beta = \frac{1}{N-\beta}\sum_{\alpha=1}^{N-\beta}\Theta(V_\alpha, V_{\alpha+\beta}) \tag{7}$$

$$\Theta(V_\alpha, V_{\alpha+\beta}) = \frac{1}{3}\{[B(V_\alpha) - B(V_\beta)]^2 + [H(V_\alpha) - H(V_\beta)]^2 + [S(V_\alpha) - S(V_\beta)]^2\} \tag{8}$$

where $B(V_\alpha)$, $H(V_\alpha)$, and $S(V_\alpha)$ are the standardized hydrophobicity value, hydrophilicity value, and side-chain mass of V_α, respectively. The dimension of the PC-PseAAC descriptor is 25.

5.2.4. Binary Profile Feature (BPF). The primary structure of a protein is composed of 20 kinds of amino acids.[40] Binary

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

profile feature (BPF)[4] is defined as the 0/1 feature code for each amino acid type. Each amino acid in each sequence is defined as a one-hot number code of length 20. For instance, the first amino acid type A can be encoded as (1,0,0,...,0), the last amino acid type Y can be encoded as (0,0,0,...,1), and so on until the binary coding vector of the entire peptide sequence can be obtained. In addition, to obtain excellent experimental results, we set the N-terminal length l of the peptide to 10 and obtained a 200-dimensional BPF feature vector. Since BPF encodes each amino acid in the sequence, BPF can usually capture more distinctive and specific information than other methods.

We not only encoded peptides based on sequence-based feature descriptors (AAC and DC) but also considered the local correlation of residues and sequence order information (PC-PseAAC). To enrich our features and make them more efficient, we added a proven feature technique, that is, we converted peptide sequences into binary map features, which can be seen as one-hot-encoding of categorical variables. Finally, we represented each ACP sequence with a 645-dimensional conjoined feature vector.

5.3. Label Propagation (LP) Network with AL. We proposed a novel active label propagation network for ACP identification. First, only randomly selected scrambled small sample data is used to train the label propagation model, and then the model predicts some unlabeled samples. Then, according to the predicted probability, the most informative samples are selected and placed in the training pool for the next model training. And so on, through a continuous selection of samples and iterative training model until the model converges.

5.4. Label Propagation with a Graph. The LP algorithm is a kind of semisupervised graph-based algorithm that constructs a graph by mapping samples to nodes, defines node similarity, and spreads label classification among similar nodes.[41] It is suitable for classification problems with few labeled data.

In the first step, we defined the labeled data with categories and unlabeled data without categories as $L = \{(x_1,y_1),(x_2,y_2),...,(x_l,y_l)\}, y_l \in \{1,2,...,\eta\}$ and $U = \{x_{l+1},x_{l+2},...,x_{l+u}\}$, respectively, where $l = u$, $n = l + u$ is the sample size and η is the number of categories (η was set to 2 in our work). Semisupervised learning refers to using L and U to predict the label $\{y_{l+1},y_{l+2},...,y_n\}$ of $\{x_{l+1},x_{l+2},...,x_n\}$.

In the second step, without the loss of generality, we used $G = (\nu,e)$ to represent an ACP network, where ν is a set of nodes representing proteins and E is the corresponding edge set between proteins. The radial basis function (RBF) is usually used to define the similarity w of two edges

$$w_{ij} = \exp\left(-\frac{\| x_i - x_j \|^2}{\alpha^2}\right) \tag{9}$$

where α is the width scale parameter, which controls the radial range of the function. The LP algorithm propagates labels through the edges between nodes. The greater the weight of the edge, the higher the similarity of nodes and the greater the probability of LP.

In the third step, the probability transition matrix is defined as follows

$$P_{ij} = p(i \to j) = \frac{W_{ij}}{\sum_{k=1}^{n} W_{ik}} \tag{10}$$

where P_{ij} is the probability of node i propagating the label to node j. Finally, all nodes update the soft label according to the probability transition matrix until the preset number of iterations or convergence. All in all, we embed the feature vector of each anticancer peptide and non-anticancer peptide into the network as a node and use the radial basis function to measure the similarity distance between the nodes as the edge weight of the network. The label propagation algorithm spreads the label in the network according to the probability conversion matrix until the node label in the network no longer changes.

5.5. AL for Data Selection. In the initial iteration of the LP algorithm, the performance was significantly improved. However, it unexpectedly started to decrease afterward.[42] An intuitive explanation for the decline in LP's accuracy is the errors accumulating over the initial iterations that overwhelm the propagation of informative labels in the subsequent iterations. This suggests coupling LP's label update with AL.

The key to LP is the data selection strategy.[20] Several strategies can usually be used to select a small set of samples, which must be annotated by a specialist, constituting the labeled part of the training set. However, the selected strategy should be able to select the samples with the most information (diversity and uncertainty) to obtain a robust classifier quickly.[43,44] In our experiments, we considered uncertain sampling and boundary data points for data selection. Uncertain sampling strategy tends to select the samples whose category is least determined by the current classifier for labeling.[43,45] When the model classifies samples, there will be samples close to the decision boundary that are similar and difficult to be classified correctly. We choose these kinds of boundary data points as information-rich samples for the model to learn. The detailed process is illustrated in Algorithm 1.

Specifically, instead of feeding all of the labeled data at once for model training, we randomly sampled N data to train and obtain the initial model (line 2). The initial model predicted all samples and provided the label probability, and we performed minimum confidence maximum label negative sorting on probabilities to form an ordered edge list (line 6). The k least confident edges at a time were obtained from the ordered list of edges and joined the training pool (line 7). We selected the k samples according to the following formula

$$x_k = \arg\max_x \left(1 - \max_{y\in\{1,0\}, l\in\{1,2,...,k-1\}} P_{lp}(y|x_l)\right) \tag{11}$$

where x_l is the label data and $P_{l_p}(y|x_l)$ is the posterior class probabilities calculated by the LP algorithm. Newly labeled data (x_k,y_k) were added to the training set (line 8) and removed from the unlabeled pool (line 9), which was no longer used in the next prediction. Then, a new LP model-based supplementary labeling data were trained (line 10). The whole process was repeated until the algorithm converged or reached the specified number of iterations (line 11). The values of N and k are discussed in Section 2.2.

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

Algorithm 1 Active learning for data selection

Input: U: unlabeled dataset;
L: labeled dataset;
Output: LP model with active strategy
initial instance selection
  select N instances from U to initial LP model M;
k instances labeling
  initialize k
  while not convergence do
    search the most informative unlabeled instance in Eq.(11)
    label x^k and add (x_k, y_k) into L:
      $L = L + \{(x_k, y_k)\}$
    remove x^k from U
    train a new LP model M based on L
    predict U by using the updated M.
  end while

In a nutshell, AL chose the most informative in the learning task and the LP algorithm used these carefully selected data to identify peptides. In addition, in the experimental part, we use the model in the sklearn learning library and applied the grid search to adjust the parameters as our comparison model.

AUTHOR INFORMATION

Corresponding Author

Xiangzheng Fu − *Department of Information Science and Technology, Hunan University, Changsha, Hunan 410000, China*; Email: fxz326@hnu.edu.cn

Authors

Lijun Cai − *Department of Information Science and Technology, Hunan University, Changsha, Hunan 410000, China*

Li Wang − *Department of Information Science and Technology, Hunan University, Changsha, Hunan 410000, China*; orcid.org/0000-0002-2801-7213

Xiangxiang Zeng − *Department of Information Science and Technology, Hunan University, Changsha, Hunan 410000, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c03132

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. L.W., X.F., and J.C. contributed to the concept and implementation. L.W., X.F., and X.Z. co-designed the experiments. L.W. and X.F. wrote the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The work was supported in part by the National Natural Science Foundation of China (61872309, 61972138, and 62002111), in part by the Fundamental Research Funds for the Central Universities (531118010355), in part by the China Postdoctoral Science Foundation (2019M662770), in part by the Hunan Provincial Natural Science Foundation of China (2020JJ4215), in part by the Scientific Research Project of Hunan Education Department (19C1788) and in part by the Changsha Municipal Natural Science Foundation (kq2014058).

REFERENCES

(1) Al-Benna, S.; Shai, Y.; Jacobsen, F.; Steinstraesser, L. Oncolytic activities of host defense peptides. *Int. J. Mol. Sci.* **2011**, *12*, 8027−8051.

(2) Kalyanaraman, B.; Joseph, J.; Kalivendi, S.; Wang, S.; Konorev, E.; Kotamraju, S. Doxorubicin-induced apoptosis: implications in cardiotoxicity. *Mol. Cell Biochem* **2002**, *234*−*235*, 119−124.

(3) Holohan, C.; Van Schaeybroeck, S.; Longley, D. B.; Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* **2013**, *13*, 714−726.

(4) Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **2018**, *34*, 4007−4016.

(5) Wu, Q.; Ke, H.; Li, D.; Wang, Q.; Fang, J.; Zhou, J. Recent Progress in Machine Learning-based Prediction of Peptide Activity for Drug Discovery. *Curr. Top. Med. Chem.* **2019**, *19*, 4−16.

(6) An, Z.; Flores-Borja, F.; Irshad, S.; Deng, J.; Ng, T. Pleiotropic Role and Bidirectional Immunomodulation of Innate Lymphoid Cells in Cancer. *Front. Immunol.* **2019**, *10*, No. 3111.

(7) Rao, B.; Zhou, C.; Zhang, G.; Su, R.; Wei, L. ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. *Briefings Bioinf.* **2020**, *21*, 1846−1855.

(8) Tesauro, D.; Accardo, A.; Diaferia, C.; Milano, V.; Guillon, J.; Ronga, L.; Rossi, F. Peptide-Based Drug-Delivery Systems in Biotechnological Applications: Recent Advances and Perspectives. *Molecules* **2019**, *24*, No. 351.

(9) Brunetti, J.; Piantini, S.; Fragai, M.; Scali, S.; Cipriani, G.; Depau, L.; Pini, A.; Falciani, C.; Menichetti, S.; Bracci, L. A New NT4 Peptide-Based Drug Delivery System for Cancer Treatment. *Molecules* **2020**, *25*, No. 1088.

(10) Esfandiari Mazandaran, K.; Mirshokraee, S. A.; Didehban, K.; Houshdar Tehrani, M. H. Design, Synthesis and Biological Evaluation of Ciprofloxacin- Peptide Conjugates as Anticancer Agents. *Iran J. Pharm. Res.* **2019**, *18*, 1823−1830.

(11) Ge, R.; Feng, G.; Jing, X.; Zhang, R.; Wang, P.; Wu, Q. EnACP: An Ensemble Learning Model for Identification of Anticancer Peptides. *Front. Genet.* **2020**, *11*, No. 760.

(12) Basith, S.; Manavalan, B.; Shin, T. H.; Lee, D. Y.; Lee, G. Evolution of Machine Learning Algorithms in the Prediction and Design of Anticancer Peptides. *Curr. Protein Pept. Sci.* **2020**, *21*, 1242−1250.

(13) Hu, Y.; Lu, Y.; Wang, S.; Zhang, M.; Qu, X.; Niu, B. Application of Machine Learning Approaches for the Design and Study of Anticancer Drugs. *Curr. Drug Targets* **2019**, *20*, 488−500.

(14) Tyagi, A.; Kapoor, P.; Kumar, R.; Chaudhary, K.; Gautam, A.; Raghava, G. P. In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* **2013**, *3*, No. 2984.

(15) Hajisharifi, Z.; Piryaiee, M.; Mohammad Beigi, M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34−40.

(16) Manavalan, B.; Basith, S.; Shin, T. H.; Choi, S.; Kim, M. O.; Lee, G. MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121−77136.

(17) Yi, H. C.; You, Z. H.; Zhou, X.; Cheng, L.; Li, X.; Jiang, T. H.; Chen, Z. H. ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation. *Mol. Ther.−Nucleic Acids* **2019**, *17*, 1−9.

(18) Liang, X.; Li, F.; Chen, J.; Li, J.; Wu, H.; Li, S.; Song, J.; Liu, Q. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. *Briefings Bioinf.* **2021**, *22*, No. 312.

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008

(19) Ahmed, S.; Muhammod, R.; Adilina, S.; Khan, Z. H.; Shatabda, S.; Dehzangi, A. ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides *BioRxiv* 2020, DOI: 10.1101/2020.09.25.313668.
(20) Bodenstedt, S.; Rivoir, D.; Jenke, A.; Wagner, M.; Breucha, M.; Muller-Stich, B.; Mees, S. T.; Weitz, J.; Speidel, S. Active learning using deep Bayesian networks for surgical workflow analysis. *Int. J. Comput. Assist. Radiol., Surg.* **2019**, *14*, 1079−1087.
(21) Zliobaite, I.; Bifet, A.; Pfahringer, B.; Holmes, G. Active learning with drifting streaming data. *IEEE Trans. Neural Netw. Learn Syst.* **2014**, *25*, 27−39.
(22) Song, M.; Yu, H.; Han, W. S. Combining active learning and semi-supervised learning techniques to extract protein interaction sentences. *BMC Bioinf.* **2011**, *12*, No. 4671.
(23) Wu, D. Pool-Based Sequential Active Learning for Regression. *IEEE Trans. Neural Netw. Learn Syst.* **2019**, *30*, 1348−1359.
(24) Carbonneau, M. A.; Granger, E.; Gagnon, G. Bag-Level Aggregation for Multiple-Instance Active Learning in Instance Classification Problems. *IEEE Trans. Neural Netw. Learn Syst.* **2019**, *30*, 1441−1451.
(25) Yu, H.; Yang, X.; Zheng, S.; Sun, C. Active Learning From Imbalanced Data: A Solution of Online Weighted Extreme Learning Machine. *IEEE Trans. Neural Netw. Learn Syst.* **2019**, *30*, 1088−1103.
(26) Kangas, J. D.; Naik, A. W.; Murphy, R. F. Efficient discovery of responses of proteins to compounds using active learning. *BMC Bioinf.* **2014**, *15*, No. 143.
(27) Doyle, S.; Monaco, J.; Feldman, M.; Tomaszewski, J.; Madabhushi, A. An active learning based classification strategy for the minority class problem: application to histopathology annotation. *BMC Bioinf.* **2011**, *12*, No. 424.
(28) Reker, D.; Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **2015**, *20*, 458−465.
(29) Hao, Z.; Lu, C.; Huang, Z.; Wang, H.; Hu, Z.; Liu, Q.; Chen, E.; Lee, C. et al. ASGN: An Active Semi-supervised Graph Neural Network for Molecular Property Prediction, *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020; pp 731−752.
(30) Lin, L.; Wang, K.; Meng, D.; Zuo, W.; Zhang, L. Active Self-Paced Learning for Cost-Effective and Progressive Face Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 7−19.
(31) Qiu, Z.; Miller, D. J.; Kesidis, G. A Maximum Entropy Framework for Semisupervised and Active Learning With Unknown and Label-Scarce Classes. *IEEE Trans. Neural Netw. Learn Syst.* **2017**, *28*, 917−933.
(32) Han, W.; Coutinho, E.; Ruan, H.; Li, H.; Schuller, B.; Yu, X.; Zhu, X. Semi-Supervised Active Learning for Sound Classification in Hybrid Learning Environments. *PLoS One* **2016**, *11*, No. e0162075.
(33) Camargo, G.; Bugatti, P. H.; Saito, P. T. M. Active semi-supervised learning for biological data classification. *PLoS One* **2020**, *15*, No. e0237428.
(34) Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K. C. iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895−16909.
(35) Yu, L.; Jing, R.; Liu, F.; Luo, J.; Li, Y. DeepACP: A Novel Computational Approach for Accurate Identification of Anticancer Peptides by Deep Learning Algorithm. *Mol. Ther.−Nucleic Acids* **2020**, *22*, 862−870.
(36) Wei, L.; Zhou, C.; Su, R.; Zou, Q. PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics* **2019**, *35*, 4272−4280.
(37) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658−9.
(38) Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T. T.; Wang, Y.; Webb, G. I.; Smith, A. I.; Daly, R. J.; Chou, K. C.; Song, J. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018**, *34*, 2499−2502.
(39) Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 246−255.
(40) Bhasin, M.; Raghava, G. P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* **2004**, *279*, 23262−23266.
(41) Hong, D.; Yokoya, N.; Chanussot, J.; Xu, J.; Zhu, X. X. Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 35−49.
(42) Kianian, S.; Khayyambashi, M. R.; Movahhedinia, N. Semantic community detection using label propagation algorithm. *J. Inf. Sci.* **2016**, *42*, 166−178.
(43) Mohamed, T. P.; Carbonell, J. G.; Ganapathiraju, M. K. Active learning for human protein-protein interaction prediction. *BMC Bioinf.* **2010**, *11*, No. S57.
(44) Zhang, X. Y.; Wang, S.; Yun, X. Bidirectional Active Learning: A Two-Way Exploration Into Unlabeled and Labeled Data Set. *IEEE Trans. Neural Netw. Learn Syst.* **2015**, *26*, 3034−3044.
(45) Liu, Y. Active learning with support vector machine applied to gene expression data for cancer classification. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1936−1941.

https://doi.org/10.1021/acsomega.1c03132
ACS Omega 2021, 6, 23998−24008