ACS OMEGA

Open Access

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

Article

Integrating Traditional Machine Learning and Deep Learning for Precision Screening of Anticancer Peptides: A Novel Approach for Efficient Drug Discovery

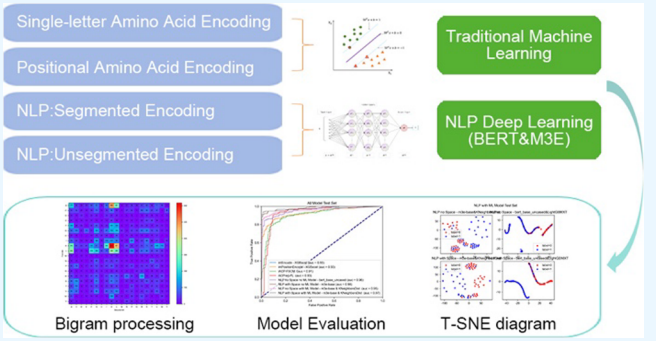
Meiqi Xu,* Jiefu Pang, Yangyang Ye, and Ziyi Zhang

Cite This: *ACS Omega* 2024, 9, 16820−16831

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The rapid and effective identification of anticancer peptides (ACPs) by computer technology provides a new perspective for cancer treatment. In the identification process of ACPs, accurate sequence encoding and effective classification models are crucial for predicting their biological activity. Traditional machine learning methods have been widely applied in sequence analysis, but deep learning provides a new approach to capture sequence complexity. In this study, a two-stage ACPs classification model was innovatively proposed. Three novel coding strategies were explored; two mainstream Natural Language Processing (NLP) models and 11 machine learning models were fused to identify ACPs, which significantly improved the prediction accuracy of ACPs. We analyzed the correlation between peptide chain amino acids and evaluated the relevant performance of the model by the ROC curve and t-SNE dimensionality reduction technique. The results indicated that the deep learning and machine learning fusion models of M3E-base and KNeighborsDist models, especially when considering the semantic information on amino acid sequences, achieved the highest average accuracy (AvgAcc) of 0.939, with an AUC value as high as 0.97. Then, in vitro cell experiments were used to verify that the two ACPs predicted by the model had antitumor efficacy. This study provides a convenient and effective method for screening ACPs. With further optimization and testing, these strategies have the potential to play an important role in drug discovery and design.



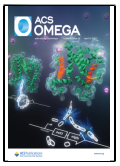
1. INTRODUCTION

Cancer is a significant global public health issue, with a rapid increase in cancer incidence and mortality rates worldwide.[1,2] It is estimated that by 2025, the annual number of new cancer cases globally will exceed 20 million.[3] Common cancer treatment modalities include surgery, chemotherapy, and radiotherapy.[4] Among these, chemotherapy is one of the primary approaches to cancer treatment. However, the major challenge faced by traditional chemotherapy is the lack of specificity of chemotherapeutic drugs, making it difficult to target tumor sites for localized treatment. This leads to collateral damage to healthy tissues and a range of complications, often necessitating discontinuation of treatment. Therefore, there is a pressing need for breakthroughs in cancer drug therapy.

Anticancer peptides (ACPs) represent a class of peptides with anticancer activity and are found in various organisms, including mammals, plants, birds, amphibians, fish, insects, and microorganisms.[5] Compared to traditional chemotherapy drugs, proteins, monoclonal antibodies and other agents, ACPs possess several distinctive characteristics, such as high specificity, ease of synthesis and modification, strong tumor penetration capabilities and diverse administration methods.[6] As an emerging anticancer therapy, ACPs have received widespread attention in recent years. In terms of clinical applications, although ACPs are still in the early stages of development, a small number of peptide drugs have successfully entered clinical trials, demonstrating their potential in treating specific cancers. With the in-depth research on ACPs and the advancement of related technologies, we have reason to believe that ACPs will play an increasingly important role in future cancer treatment and bring new hope to cancer patients.

Ideally, anticancer peptides can selectively target cancer cells without harming normal tissue cells, inducing cancer cell death by altering cell membrane permeability, or interacting with intrinsic targets within cancer cells. The mechanisms of action of ACPs from different sources and their modified peptides have become a recent focus of research in anticancer drug

Received: February 12, 2024
Revised: March 3, 2024
Accepted: March 22, 2024
Published: April 1, 2024



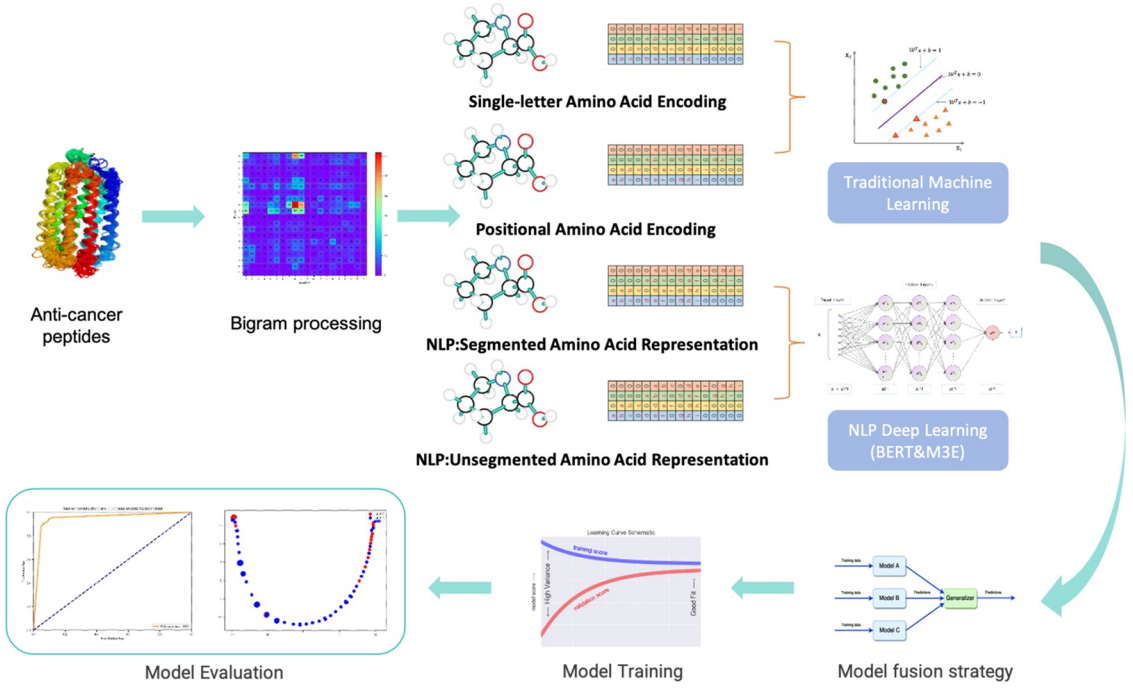


Figure 1. Workflow of this study.

development.[7−9] Some ACP-based drugs have already entered clinical trials or been approved for use.[10] Therefore, the rapid identification of potential ACPs holds significant importance for the development of cancer treatments. However, traditional methods of identifying ACPs through wet laboratory experiments are relatively costly and time-consuming. Computational techniques in the field of bioinformatics provide a solution to the rapid and accurate identification of ACPs. Among various computational methods, machine learning has emerged as a promising approach to efficiently identifying ACPs.

Traditional machine learning methods for ACP identification involve manually designing features for classifying protein sequences, followed by the use of classification models such as Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Extra Trees and Gradient Boosting Trees.[11−14] However, relying solely on handcrafted features may not fully capture the complexity and diversity of data features.[15,16] Deep learning methods, known for their ability to efficiently handle unstructured data, have increasingly been applied in ACP identification.[17] Yi et al. introduced a deep learning Long Short-Term Memory (LSTM) neural network model, ACP-DL, which automatically learned how to identify anticancer peptides and nonanticancer peptides by integrating binary profile features and a simplified amino acid alphabet sparse matrix.[18] However, this method still fundamentally used handcrafted features as inputs, limiting the neural network's ability to acquire raw information. Yang et al. proposed the CACPP model, which employed a convolutional neural network (CNN) to extract high-potential features from peptide sequences and a contrastive learning module to learn more distinguishable feature representations in a deep learning framework.[19] However, due to the restricted receptive field of the CNN network, it was challenging to capture the interactions between amino acids that are distant from each other in long peptide chains. Sun et al. proposed an ACP-BC model, which uses a bidirectional long short-term memory network (BiLSTM) to extract features from the original sequence and obtains deep abstract features through a bidirectional encoder representation converter (BERT), achieving an accuracy of 87%.[20] However, handcrafted features also limited the neural network's ability to acquire raw information, and using chemical formulas as inputs for the BERT network made it prone to truncation when peptide chains were too long. Currently, it appears that neural networks commonly used in computer vision, such as CNNs, and sequence-based neural networks, such as LSTMs, are more prevalent in ACP identification models. However, we believe that models related to Natural Language Processing (NLP) are more suitable for ACP screening.

In this study, we employed four different feature representation and encoding methods, and trained and compared various deep learning and machine learning models to screen peptide sequences with anticancer potential. First, we performed bigram processing on amino acid sequences to generate overlapping matrices and conducted relevant statistical analysis and t-SNE dimensionality reduction visualization. This aided in exploring the co-occurrence relationships and correlations between amino acids, providing support for the design and prediction of peptide sequences with anticancer potential. In the feature engineering phase, we utilized three encoding methods: (1) encoding considering only the amino acid letter information, (2) encoding considering both amino acid letter information and amino acid position information, and (3) encoding the amino acid sequence as if it were a sentence. In the third encoding method, we used two different NLP sequence representations, with or without splitting amino acids, and performed transfer training using two NLP models, BERT and M3E. We employed a greedy soup strategy to enhance network generalization and trained various machine learning models on top of deep learning. The results for each method included peptide sequence classification as well as performance evaluation using tools such as ROC curves and t-SNE dimensionality reduction plots. We identified the model combination of the m3e-base and KNeighborsDist with the

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

highest average accuracy (ValACC and TestACC) of 93.85%. Using this model, we screened peptide sequences effectively validated by both the original NLP model and the comprehensive model, followed by synthesis and in vitro antitumor efficacy validation. The results indicated that the peptides validated by the model demonstrated significant anticancer cell cytotoxicity. Overall, our innovative approach combining NLP language models and machine learning models for ACP screening achieved high accuracy and ease of implementation compared with traditional screening methods. Through in vitro experiments, we confirmed that the peptides selected by our model possess effective anticancer activity, which could aid researchers in discovering more potential ACPs and promoting the development of ACP-based drugs. The workflow of this study was illustrated in Figure 1.

2. MATERIALS AND METHODS

2.1. Data Collection. The data for this study were collected from various publicly available anticancer peptide databases, including APD3 (https://aps.unmc.edu/database/anti), BioPepDB (http://bis.zju.edu.cn/biopepdbr/index.php?p=search&field=category&query=anticancer), DPL (http://www.peptide-ligand.cn/search/?csrfmiddlewaretoken=PSqYxvTcUbmCHIAOCjLDJa0tzZky9MoQ6YR9NrAVsHHqeUB6uBkdMNyrmIJ1o2Zf&q0=&q1=&q2=&q4=Anticancer+&q3=&q5=&submit=Search), PlantPepDB (http://14.139.61.8/PlantPepDB/pages/browse_result.php), and FeptideDB (http://www4g.biotec.or.th/FeptideDB/peptide_search.php). These diverse databases provided comprehensive data support by offering a wide range of anticancer peptide sequence information. To ensure data integrity, duplicate peptides within individual databases and between databases were removed, and the anticancer activity of each peptide was rigorously validated. The integrated database consists of 859 anticancer peptides. Additionally, to balance the data set, an equal number of random peptides were generated using a randomization process to serve as control samples for training and testing.

2.2. Data Preprocessing. Data preprocessing played a critical role in this study to ensure the quality and suitability of the amino acid sequence data. Strategies employed included identifying inconsistent or unexpected format data within sequences and correcting or removing them to ensure data set consistency and reliability. Furthermore, techniques such as imputation, deletion, or interpolation were applied to address missing values or blank data, ensuring data completeness and usability. In addition, noise data were also processed during data cleaning using filtering or other denoising methods to reduce random interference in the data, ensuring the stability and reliability of subsequent analyses.

After data cleaning, the Bigram method for processing amino acid sequences was a key step in this study.[21] By applying Bigram processing to amino acid sequences, we generated an overlapping matrix. This matrix recorded the pairwise occurrences of amino acids in the sequences and, through statistical analysis methods, explored co-occurrence relationships and correlations between amino acids. Through the analysis of these relationships, we identified potential patterns and feature sequences with anticancer potential.

2.3. Peptide Sequence Feature Extraction and Representation. In this study, the feature extraction methods covered three novel encoding methods aiming to better capture the features of amino acid sequences. First, the first method considered only the encoding of the amino acid letters themselves, mapping different amino acids to specific symbols or numbers. Second, the second method comprehensively considered the letter and position information on amino acids for a more comprehensive representation of amino acid sequence features. Lastly, the third method viewed amino acid sequences as natural language texts, employing two different NLP sequence representation methods: one split the amino acid sequence into individual amino acids as words and the other treated the entire amino acid sequence as a sentence. Both methods used models from NLP (e.g., BERT and M3E) to represent amino acid sequences in vectorized form, facilitating understanding and processing by deep learning models.

2.4. Screening Model Construction. *2.4.1. Construction of Screening Models Based on Machine Learning.* In this study, for screening anticancer peptides, the research team chose a variety of machine learning algorithms for comparative analysis. These algorithms included LightGBMLarge, LightGBMXT, XGBoost, LightGBM, CatBoost, KNeighborsDist, KNeighborsUnif, RandomForestEntr, RandomForestGini, ExtraTreesEntr and ExtraTreesGini.

For model training and optimization, exhaustive parameter adjustments and optimizations were conducted for the machine learning models. For gradient boosting tree models such as LightGBM, XGBoost, and CatBoost, key parameters such as learning rate, tree depth, and number of leaf nodes were adjusted to improve model generalization and predictive performance. In KNN models, parameters such as the number of neighbors and distance measurement methods were adjusted. For ensemble learning methods such as Random Forests and Extra Trees, parameters such as the number of trees and maximum number of features were optimized to achieve the best model effect. During the training process, methods such as grid search were used to further enhance model stability and performance.

2.4.2. Construction of Screening Models Based on Deep Learning. Specifically, two main deep learning models, BERT (Bidirectional Encoder Representations from Transformers) and M3E (Multiscale 3D Epitope Prediction), were chosen and explored in this study. BERT, a milestone in the NLP field with a Transformer structure, excels in capturing sequence features and context. M3E, a multitask model, has been enhanced for text classification tasks, allowing it to process multiscale features between sequences, which is advantageous for protein function prediction.

Moreover, in addition to the deep learning models, machine learning models were also incorporated to enhance the model performance. Specifically, comparative experiments were conducted in the BERT and M3E models, including using only NLP models and combining them with various machine learning models. These machine learning models include the ones listed above, such as LightGBMLarge, XGBoost, KNeighbors, etc., used to further enhance predictive performance on top of deep learning models.

During the model training and optimization process, this study conducted exhaustive tuning for different deep learning models and strategies. For BERT and M3E models, key hyperparameters such as learning rate, number of layers, and number of hidden units were adjusted, and techniques like pretraining and fine-tuning were used to improve model performance. In the model combination stage, parameters of different machine learning models were adjusted, and

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

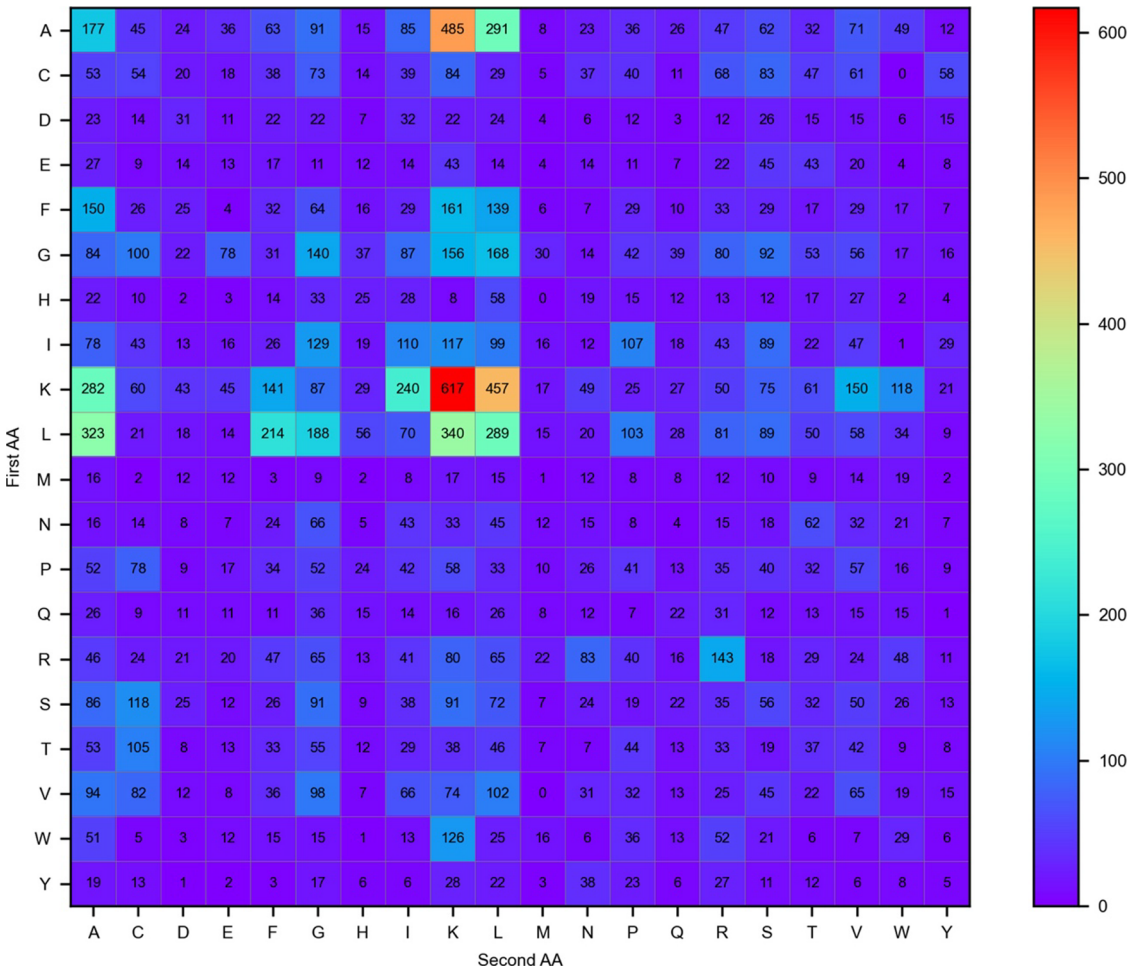


Figure 2. Overlapping matrix visualization.

optimization methods such as grid search were used. Throughout the training process, appropriate loss functions and evaluation metrics were used to obtain more robust models with stronger generalization capabilities.

2.5. Experimental Design. In the experimental design of this study, a strict validation and test data set division approach was adopted (60% of the data for training, 20% for validation and tuning model parameters, and the final 20% for the ultimate model testing and evaluation). For the optimization of model parameters, methods like grid search were used to determine the best combination of hyperparameters, thereby enhancing the model's generalizability. Regarding computational equipment and configuration, the experiment utilized high-performance computing servers equipped with high-memory GPUs (NVIDIA GeForce RTX 2070 graphics cards) to accelerate the model training and optimization process. The experimental environment was based on the Python programming language, employing relevant deep learning frameworks (such as PyTorch) and machine learning libraries (such as Scikit-learn), ensuring stability and efficiency of the experimental platform.

2.6. Model Evaluation. In the model evaluation process of this study, the performance metrics used covered multiple aspects, including Accuracy, Recall, Precision, F1 Score, and Area Under the Curve (AUC). These metrics were capable of comprehensively assessing the model's performance in various aspects.

Accuracy intuitively reflects the overall predictive accuracy of the model, that is, the proportion of correctly predicted samples, and is a key indicator of the model's basic performance. Recall measures the proportion of anticancer peptides identified by the model out of the total actual anticancer peptides, focusing on the model's coverage capability. Precision assesses the proportion of samples correctly identified as anticancer peptides by the model out of the total samples predicted as anticancer peptides, reflecting the model's accuracy. The F1 Score, which is the harmonic mean of precision and recall, provides an evaluation of comprehensive performance, particularly suitable for imbalanced data sets. Lastly, the AUC value measures the model's overall performance at different thresholds, serving as an important indicator of the model's predictive capacity. Together, these evaluation metrics enable us to understand and assess the model's performance from multiple dimensions comprehensively, ensuring the model's effectiveness and reliability in practical applications.

2.7. In Vitro Antitumor Efficacy of Synthetic Peptides. The model with the best effect was selected to predict two peptides with a high antitumor efficacy for synthesis. The two peptide sequences were purchased from Jill Biochemical Co., Ltd. (Shanghai, China) and synthesized by the synthesis method of fluorenylmethyloxy carbonyl chloride protective amino acids. The purity of the synthetic peptide was detected by high-performance liquid chromatography (HPLC) to be

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

more than 95%. The in vitro cytotoxicity of peptides was evaluated in 4T1 cells (mouse breast cancer cells) using the sulfonyl rhodamine B (SRB) assay (Sigma-Aldrich, St. Louis, MO, USA).[22,23] 4T1 cells (6×10^3 cells/well) were seeded in 96-well plates, incubated for 24 h, and then treated with peptides at 37 °C for 24 h. The cell viability was determined using SRB, which allowed quantification of the living cells by measuring absorbance at 540 nm with a 96-well plate reader (model 680; Bio-Rad Laboratories Inc., Hercules, CA, USA). The half-maximal inhibitory concentration (IC50) was calculated according to the dose−effect curves using GraphPad Prism 8 software.

3. RESULTS

3.1. Distribution Statistics Display. The activity of ACPs is influenced by the composition of amino acids and their structure. Previous studies mostly considered the frequency of individual amino acids in ACPs.[24] However, we believed that the co-occurrence relationships and correlations between amino acids are more meaningful in the search for and exploration of peptides with anticancer activity. Therefore, we processed the amino acid sequences through bigram processing, generated overlapping matrices, and conducted a statistical analysis of the results.

As shown in the context matrix plot (Figure 2), lysine (Lys, K), alanine (Ala, A), and leucine (Leu, L) had the highest frequency in ACPs. This may be attributed to the positively charged nature of these amino acids or their hydrophobic properties. When ACPs interact with cancer cell membranes, the positively charged hydrophilic regions effectively bind to the negatively charged surface of cancer cell membranes through electrostatic adsorption. Simultaneously, the hydrophobic regions bind to the membrane lipids. This laid the foundation for the selective action of ACPs on cancer cells, and the results were consistent with the findings reported in the literature.[25] It was noteworthy that the top three amino acid pairs in terms of occurrence frequency are K-K, A-K, and K-L, with frequencies of 617, 485, and 457, respectively (Table 1). Identifying amino acid pairs with higher occurrence frequencies was more helpful in designing amino acid sequences with anticancer activity based on contextual relationships, which could not be achieved by only counting individual amino acids.

3.2. Machine Learning Model Performance Evaluation Results. *3.2.1. Results of Encoding Based on Amino Acid Letters Only.* We first explored the impact of encoding based on amino acid letters on the performance of machine learning models. This encoding strategy maps different amino acids to a set of unique symbols or numbers, serving as the basis for training and predicting machine learning models. To comprehensively assess the effectiveness of this encoding method, we employed a series of advanced machine learning algorithms, including but not limited to LightGBMLarge, XGBoost, CatBoost, etc. Each algorithm was evaluated on multiple dimensions including precision (Pre), recall (Rec), F1 score, and accuracy (ACC). The specific results are shown in Table 2. In the comparative analysis, we observed that the XGBoost model achieved a high score of 0.873 in average ACC, and its precision, recall, and F1 scores were also impressive at 0.909, 0.851, and 0.879, respectively. Furthermore, by conducting ROC curve analysis for the top 6 models ranked by ACC (Figure 3a), both XGBoost and LightGBMLarge models demonstrated outstanding performance with AUC values reaching 0.93, highlighting their excellence in classification tasks.

Table 1. Number of Occurrences of Amino Acid-Amino Acid Pairs

amino acid-amino acid pairs	number of occurrences
K-K	617
A-K	485
K-L	457
L-K	340
L-A	323
A-L	291
L-L	289
K-A	282
K-I	240
L-F	214
L-G	188
A-A	177
G-L	168
F-K	161
G-K	156
K-V	150
F-A	150
R-R	143
K-F	141
G-G	140

We further performed visual analysis of features using the t-SNE dimensionality reduction technique. Despite the simplicity and intuitiveness of this encoding method, the results in Figure 3b revealed a challenge. There seemed to be a difficulty in distinguishing between ACPs and non-ACPs using features extracted by this encoding method. This might be attributed to the fact that this encoding strategy overlooks spatial positions and contextual information in amino acid sequences, which often play crucial roles in biological sequence analysis.

3.2.2. Results of Encoding Considering Both Amino Acid Letter and Position Information. Next, we delved into a more complex encoding strategy that not only included the letter information on amino acids but also integrated their positional information in the sequence to achieve a comprehensive representation of amino acid sequence features. By introducing positional information, we aimed to capture the spatial structural characteristics of amino acid sequences, thereby enhancing the model's classification ability for ACPs and non-ACPs. As shown in Table 3, we evaluated various machine learning models, including but not limited to XGBoost, LightGBM, etc. The results indicated that when positional information is introduced, the top-performing models remained comparable to the first encoding method across all metrics. XGBoost and LightGBMLarge models continued to exhibit the best performance. Furthermore, we conducted a comparative analysis of model performance through ROC curves (Figure 4a), and the results corroborated the numerical indicators in the table, demonstrating that the overall performance of the models was maintained.

To further analyze the impact of encoding on classification performance, we applied the t-SNE technique for dimensionality reduction and visualized the model's classification results (Figure 4b). Despite the introduction of positional information, there was an improvement in the separation of the models in distinguishing ACPs from non-ACPs. However, the results

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

Table 2. Machine Learning Model Results of Encoding Based on Amino Acid Letters Only

ML model	ValAcc	TestAcc	TestPre	TestRec	TestF1	AvgAcc
XGBoost	0.868	0.877	0.909	0.851	0.879	0.873
LightGBMLarge	0.871	0.868	0.897	0.846	0.871	0.870
LightGBM	0.883	0.853	0.871	0.846	0.858	0.868
CatBoost	0.868	0.835	0.857	0.823	0.840	0.852
LightGBMXT	0.856	0.841	0.855	0.840	0.847	0.849
RandomForestEntr	0.856	0.835	0.880	0.794	0.835	0.846
RandomForestGini	0.838	0.829	0.864	0.800	0.831	0.834
ExtraTreesEntr	0.817	0.826	0.855	0.806	0.829	0.822
ExtraTreesGini	0.814	0.820	0.849	0.800	0.824	0.817
KNeighborsDist	0.643	0.707	0.668	0.874	0.757	0.675
KNeighborsUnif	0.637	0.701	0.665	0.863	0.751	0.669

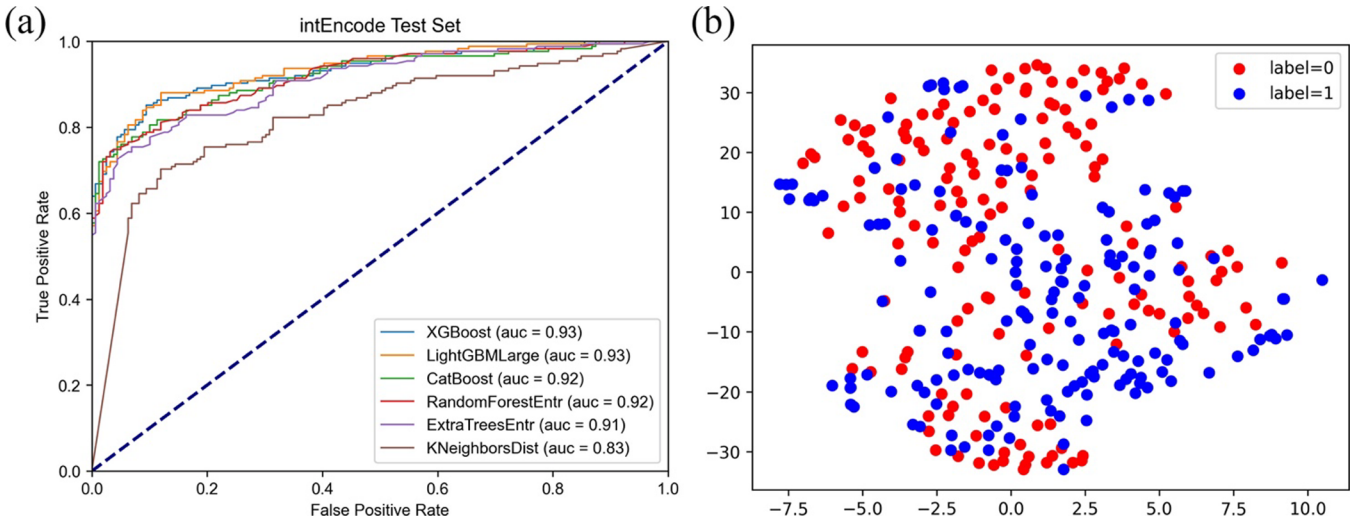


Figure 3. Visual analysis of encoding based on amino acid letters only. (a) ROC curves and AUC values. ROC: receiver operating characteristic; AUC: area under the ROC curve. (b) T-SNE dimensionality reduction diagram.

Table 3. Machine Learning Model Results of Encoding Considering Both Amino Acid Letter and Position Information

ML model	ValAcc	TestAcc	TestPre	TestRec	TestF1	AvgAcc
XGBoost	0.868	0.877	0.909	0.851	0.879	0.873
LightGBMLarge	0.871	0.868	0.897	0.846	0.871	0.870
LightGBM	0.883	0.853	0.871	0.846	0.858	0.868
CatBoost	0.868	0.835	0.857	0.823	0.840	0.852
LightGBMXT	0.856	0.841	0.855	0.840	0.847	0.849
RandomForestEntr	0.856	0.835	0.880	0.794	0.835	0.846
RandomForestGini	0.844	0.832	0.865	0.806	0.834	0.838
ExtraTreesGini	0.808	0.832	0.848	0.829	0.838	0.820
ExtraTreesEntr	0.799	0.832	0.848	0.829	0.838	0.816
KNeighborsDist	0.682	0.704	0.692	0.783	0.735	0.693
KNeighborsUnif	0.670	0.689	0.680	0.766	0.720	0.679

suggested that this improvement was not significant, and the models still faced challenges in clearly distinguishing between the two sequence classes. This might indicate that simply adding positional information had limitations in enhancing the model's ability to recognize ACPs, or it might suggest the need for further optimization of feature expression and model structure to fully leverage positional information. In conclusion, while the second encoding method provided a more complex and comprehensive feature representation by considering both amino acid letters and positional information, its practical application did not significantly improve model performance.

3.3. Deep Learning Model Performance Evaluation Results. *3.3.1. Results without Adding Machine Learning Models.* In an effort to improve the results of traditional machine learning models, we experimented with two NLP deep learning models—BERT and M3E. Simultaneously, we employed a third encoding method that treats amino acid sequences as natural language text and utilized two different NLP sequence representation approaches: one treating the entire amino acid sequence as a single sentence and the other splitting the amino acid sequence into individual amino acids as words. This was done to vectorize the amino acid sequences for understanding and processing by deep learning models. As

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

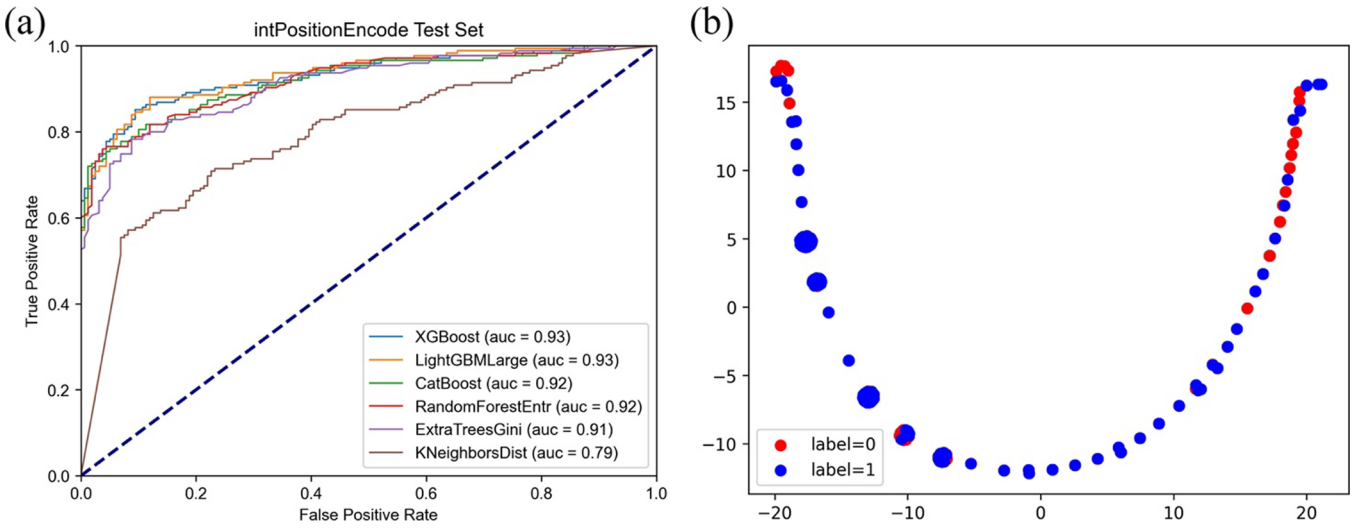


Figure 4. Visual analysis of encoding considering both amino acid letter and position information. (a) ROC curves and AUC values. (b) T-SNE dimensionality reduction diagram.

Table 4. Results of Deep Learning Models with Two Different NLP Sequence Representations

	DL model	ValAcc	TestAcc	TestPre	TestRec	TestF1	AvgAcc
NLP no space	m3e-base	0.904	0.889	0.879	0.914	0.896	0.897
	bert-base-uncased	0.904	0.904	0.961	0.851	0.903	0.904
NLP with space	m3e-base	0.904	0.940	0.953	0.931	0.942	0.922
	bert-base-uncased	0.910	0.919	0.925	0.920	0.923	0.915

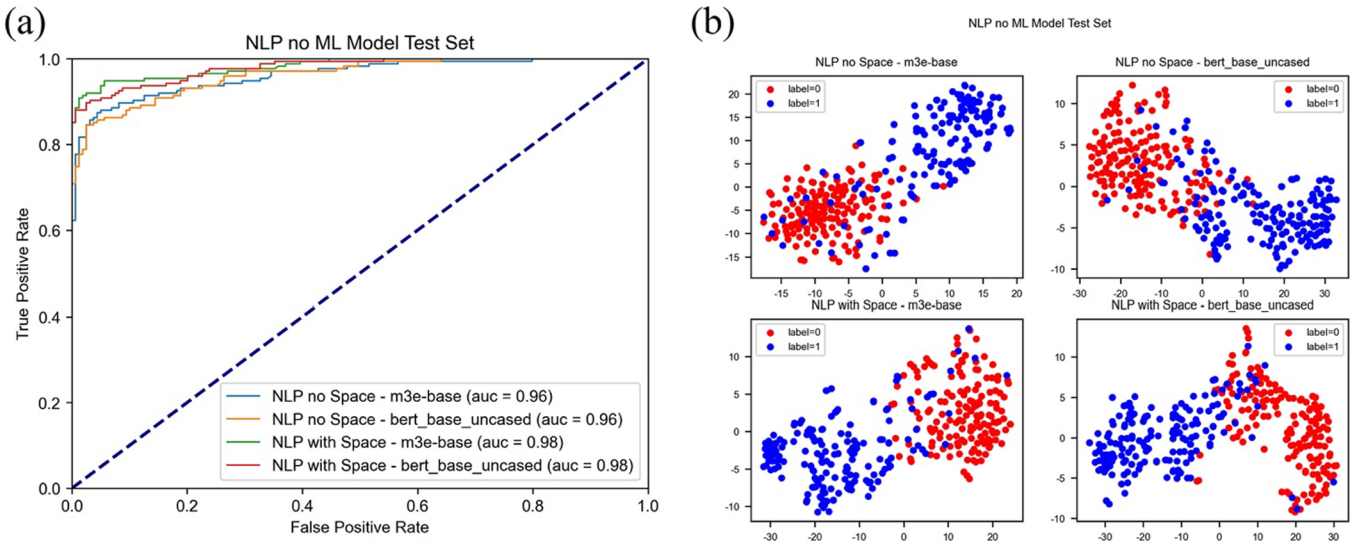


Figure 5. Visual analysis of deep learning models. (a) ROC curves and AUC values. (b) T-SNE dimensionality reduction diagram.

shown in Table 4, in the first NLP sequence representation method (treating the entire amino acid sequence as a single sentence), the BERT model performed better with test set ACC, Pre, Rec, and F1 scores of 0.904, 0.961, 0.851, and 0.903, respectively. The average ACC for the test and validation sets was 0.904. In the second NLP sequence representation method (splitting the amino acid sequence into individual amino acids as words), the M3E model showed superior performance, surpassing the optimal results of the first representation method. The test set ACC, Pre, Rec, and F1 scores reached 0.919, 0.925, 0.920, and 0.923, respectively, with the average ACC for the test and validation sets reaching 0.915. The ROC curve demonstrated an AUC value of 0.96 for the first representation method and a remarkable AUC value of 0.98 for the second representation method (Figure 5a). The t-SNE dimensionality reduction plot also indicated that the second representation method combined with the M3E model had better discriminability for the ACPs (Figure 5b). This aligns with reality as splitting amino acids in the sequence into individual words with a natural order is more consistent with the natural distribution of amino acids in the sequence. Therefore, transforming amino acid sequences into vector representations with semantic information can better capture advanced features and relationships in the sequence, leading to a noticeable improvement in model accuracy.

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

Table 5. Results of Fusion Models of Deep Learning and Machine Learning

	NLP model	AvgAcc of best ML model	ValAcc	TestAcc	TestPre	TestRec	TestF1	AvgAcc
NLP no space	m3e-base	KNeighborsDist	0.907	0.922	0.963	0.886	0.923	0.915
	bert-base-uncased	LightGBMXT	0.901	0.901	0.928	0.880	0.903	0.901
NLP with space	m3e-base	KNeighborsDist	0.919	0.958	0.976	0.943	0.959	0.939
	bert-base-uncased	LightGBMXT	0.913	0.931	0.958	0.909	0.933	0.922

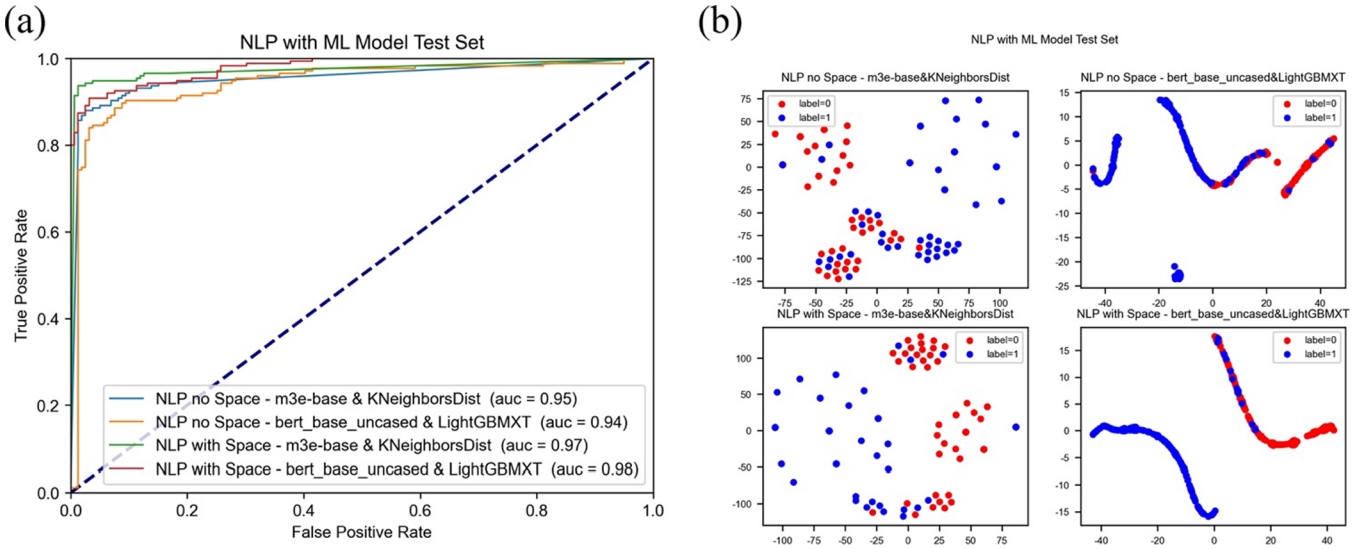


Figure 6. Visual analysis of fusion models of deep learning and machine learning. (a) ROC curves and AUC values. (b) T-SNE dimensionality reduction diagram.

3.3.2. Results of Fusion Models of Machine Learning and Deep Learning. In this study, to enhance the accuracy of deep learning models, we not only independently utilized NLP deep learning models, but also attempted to combine them with traditional machine learning models using the Greedy Soup strategy to achieve a superior performance. Through this fusion approach, we could leverage the ability of deep learning models to extract complex features from amino acid sequences and enhance the final classification accuracy by using machine learning models. We selected the model with the highest ACC from numerous combinations for detailed evaluation, and the results were listed in Table 5. The findings indicated that the overall performance of the model improved after incorporation of machine learning models. In particular, the combination of the M3E-base and KNeighborsDist models exhibited the highest accuracy in all tests, reaching an impressive 93.9%.

On the other hand, we visually presented the experimental results of this fusion strategy. The ROC curve graph clearly illustrated the classification performance of the fusion model (Figure 6a). We observed that when not considering the semantic information on amino acids, the combination of M3E-base and KNeighborsDist achieved an AUC value of 0.95, demonstrating excellent classification capability. In the presence of semantic information, the combination of M3E-base and KNeighborsDist increased the AUC value to 0.97, while the combination of bert_base_uncased and LightGBMXT reached an AUC value of 0.98. This highlighted the significance of semantic information in enhancing model performance.

Furthermore, the t-SNE dimensionality reduction plot revealed the separation of the model in handling different categories of data (Figure 6b). Comparing the dimensionality reduction results of different model combinations, we found that considering spatial information significantly improved the model's ability to differentiate between ACPs (label = 1) and non-ACPs (label = 0). In particular, the combination of bert_base_uncased and LightGBMXT exhibited a higher category separation in the dimensionality reduction plot, corroborating its high AUC value in the ROC curve graph.

These results not only confirmed the effectiveness of NLP deep learning models in handling amino acid sequences but also indicated that traditional machine learning models could effectively enhance the performance of deep learning models, especially when incorporating semantic information into the sequence. Through this hybrid model, we could more accurately identify and classify ACPs, thereby applying them more effectively in drug design and related fields of bioinformatics.

3.3.3. Comprehensive Results and Performance Comparison. This study conducted a comprehensive comparison of the impact of different encoding methods and model combinations on the task of identifying ACPs. By contrasting various encoding strategies and deep learning (DL) models with or without the inclusion of machine learning (ML) models, our determined the optimal model configuration for achieving high-precision recognition of ACPs. In addition, we compared the performance of our models with several currently known excellent ACP predictors. In Table 6, we compared six different encoding and model combination strategies, recording their average accuracy (AvgAcc). XGBoost models based on integer encoding (intEncode) and integer encoding considering amino acid position information (intPositionEncode) both demonstrated an AvgAcc of 0.873, providing a stable baseline performance. However, when transitioning to NLP deep learning models, we noted that bert-base-uncased achieved an AvgAcc of 0.904 without

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

Table 6. Highest AvgAcc Models under Different Encoding Methods and Comparisons with Other ACP Predictors

encoding method	model with the highest AvgAcc	AvgAcc
intEncode	XGBoost	0.873
intPositionEncode	XGBoost	0.873
NLP no space (no ML model)	bert-base-uncased	0.904
NLP with space (no ML model)	m3e-base	0.922
NLP no space (with ML model)	m3e-base + KNeighborsDist	0.915
NLP with space (with ML model)	m3e-base + KNeighborsDist	0.939
iACP-FSCM		0.857
ACPred-FL		0.898

semantic information and without combining ML models. Furthermore, the m3e-base model, when considering semantic information and not combining ML models, reached an AvgAcc of 0.922. When combining DL models with ML models, performance showed a significant improvement. In particular, when not considering semantic information, the combination of the m3e-base and KNeighborsDist increased AvgAcc to 0.915. In the presence of semantic information, the same combination achieved the highest AvgAcc value in this study, reaching 0.939.

The ROC curve graph in Figure 7 provided an intuitive way to compare performance. From the graph, we observed that the combination of m3e-base and KNeighborsDist, which incorporated semantic information, demonstrated the most outstanding classification ability among all of the models, with an AUC value of 0.97. This further emphasized the importance of semantic information in enhancing model accuracy and the potential of ML models to effectively boost the performance of DL models.

Then, we selected two currently known excellent ACP predictors, iACP-FSCM[26] and ACPred-FL,[27] tested them with our data set, and compared their ROC curves and AccAvg values with our models. The performance comparison results showed that our models were significantly higher than the iACP-FSCM and ACPred-FL models in terms of the AUC value and AccAvg value. It proved that our models were advanced compared to existing ACP predictors.

3.4. Model Prediction and In Vitro Antitumor Efficacy of Peptides. We randomly selected 741 peptides as test peptides in the database and predicted them using the model m3e-base+KNeighborsDist with the highest comprehensive evaluation. Two peptides with potentially high anticancer activity were predicted, namely, KEWLE and KRLAFA. Mass spectrometry results showed that the two peptides were successfully synthesized, and HPLC results confirmed that the purity of both peptides was above 95% (Figures S1−S4). Mouse breast cancer cell 4T1 cells were selected to validate the in vitro antitumor efficacy of the two peptides. The experimental results are shown in Table 7, and the IC50 values of the two peptides were, respectively, 47.46 ± 12.14 and 39.99 ± 9.744 μM, and both had in vitro antitumor efficacy.

Table 7. IC50 Values (μM) of Peptides in 4T1 Cell Lines

peptide sequences	molecular weight	IC50 (μM)
KEWLE	703.78	47.46 ± 12.14
KRLAFA	704.86	39.99 ± 9.744

4. DISCUSSION

This study aimed to construct a precise and automated model for the screening of ACPs using machine learning methods. Initially, three different feature representation and encoding methods were employed, with the third encoding method combining two different NLP sequence representations, utilizing NLP models such as BERT and M3E, and introducing a combination strategy of machine learning models. By integration of various machine learning and deep learning models, a model combination with higher accuracy was ultimately selected. In vitro anticancer efficacy validation was performed, demonstrating that the peptides validated by the model exhibited potent anticancer cell toxicity. Overall, this

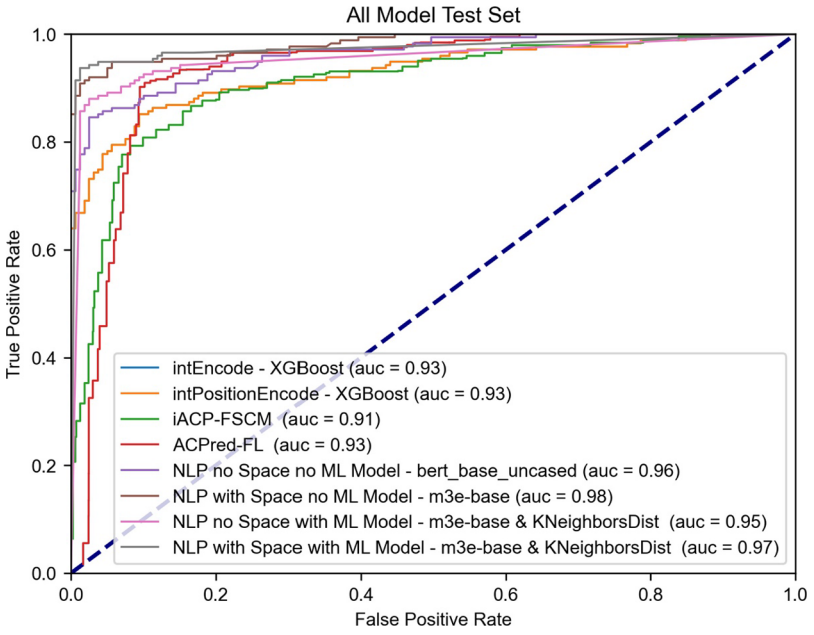


Figure 7. ROC curves and AUC values of the highest AvgAcc models and comparisons with other ACP predictors.

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

study innovatively incorporated NLP into the screening of anticancer peptides, achieving higher and more achievable accuracy compared to traditional methods and confirming the practical and effective anticancer activity of the selected peptides.

The machine learning methods employed in this study, such as LightGBM, XGBoost, etc., demonstrated good generalization performance in terms of model efficiency, especially when dealing with high-dimensional features and large-scale data, exhibiting high efficiency and prediction accuracy. These machine learning models effectively screened anticancer peptides through traditional feature engineering and model fusion strategies, confirming their reliability in peptide sequence analysis. In contrast, deep learning methods such as BERT, M3E, etc., showed advantages in capturing sequence features and contextual information. These models could better utilize the sequential information and global context in the sequences, enhancing the model's predictive capabilities through learning advanced features. Particularly, the BERT model, as an advanced model in the natural language processing field, successfully applied the Transformer structure to learn representations of protein sequences, offering new possibilities for the screening of anticancer peptides. In summary, traditional machine learning models have advantages in efficiency, while deep learning models excel in learning and representing sequence features. Therefore, combining machine learning and deep learning methods may become a new trend in the future for screening anticancer peptides. It is worth noting that compared with the studies by Wei et al.[27] and Charoenkwan et al.,[26] this study focuses on predicting the activity of ACP using a combination of deep learning and efficient feature expression, achieving the highest prediction accuracy of 0.939. The above comparative studies construct effective sequence feature expressions and achieve prediction through traditional machine learning, and the prediction accuracy is slightly lower than that in this study.

It was noteworthy that this experiment demonstrated significant practical value in the screening of anticancer peptides. By application of a combination of machine learning and deep learning models, the efficiency of anticancer peptide screening was improved. Traditional methods of anticancer peptide screening typically required extensive experimentation and human resources. The model constructed in this study could rapidly and accurately predict peptide sequences with potential anticancer activity, significantly shortening the screening cycle. This precise screening method effectively reduced the cost of developing new anticancer drugs, accelerating the drug development process. This is also in line with the future development trend in this field using advanced computational methods to discover functional peptides to aid drug screening.[28]

From a longer-term perspective, ACPs have emerged as a promising alternative due to their specificity, lower toxicity, and ability to circumvent drug resistance mechanisms that often limit the effectiveness of traditional therapies. Traditional chemotherapy indiscriminately targets rapidly dividing cells, leading to significant side effects, whereas ACPs offer a more targeted approach, potentially leading to more effective and less harmful treatments. Moreover, advancements in peptide engineering and drug delivery systems have significantly enhanced the stability, bioavailability, and tumor-targeting capabilities of ACPs, making them a viable option in the arsenal against cancer. However, despite these advantages, the clinical application of ACPs still faces several challenges such as their rapid degradation in the bloodstream, potential immunogenicity, and the complexity of large-scale synthesis. These issues necessitate further research and innovation to overcome. Additionally, while ACPs show promise in preclinical studies, their efficacy and safety in human trials must be rigorously tested. The possibility of replacing traditional anticancer drugs with peptides depends not only on overcoming these technical hurdles but also on demonstrating superior clinical outcomes.

While this study has made significant progress in anticancer peptide screening, there are limitations that need to be considered. First, the quality and quantity of the data set have a crucial impact on the training and prediction of machine learning and deep learning models, and the data set used in this study may have limitations. The size and coverage of the data set may influence the generalization ability and predictive performance of the model. Therefore, a more extensive and representative data set could further enhance the robustness of the model. Second, the choice of feature representation and encoding methods has a significant impact on the performance of the model. Although the feature engineering methods used in this study are diverse, there may still be a potential optimization space. Further exploration of more effective feature representation and encoding methods may improve the model's ability to abstractly represent peptide sequences, thereby enhancing predictive performance. Lastly, despite the in vitro experiments validating the anticancer activity of the peptides screened by the model, more experimental data and further clinical trials are needed to verify the feasibility and effectiveness of their clinical application. There is a certain difference between laboratory conditions and actual treatment environments, necessitating more clinical data to support the experimental results obtained in this study.

5. CONCLUSIONS

This study successfully integrated different encoding methods with multiple machine learning approaches to construct a precise screening model for anticancer peptides. Through rigorous model training and comparison, the study accurately screened peptide sequences with anticancer potential and confirmed the effectiveness of the model's predictions in in vitro experiments. Particularly, the introduction of NLP language models into the screening of anticancer peptides in this study, compared with traditional methods, significantly improved both the accuracy and efficiency of screening. This opened up new possibilities and avenues for the discovery and research of anticancer drugs. However, despite the relatively outstanding results achieved in this study, there were still some suggestions for improvement to consider. In future research, we will continue to explore and incorporate the latest technologies and theoretical advancements to continuously enhance the performance and practicality of the model, promoting the development and application of anticancer drugs.

■ ASSOCIATED CONTENT

Data Availability Statement

All code used in data analysis and preparation of the manuscript, alongside a description of necessary steps for reproducing results, can be found in a GitHub repository: https://github.com/Keda0411/acps.

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c01374.

Characterization of synthetic peptides. Mass spectrometry of KEWLE and KRLAFA and HPLC of KEWLE and KRLAFA (PDF)

AUTHOR INFORMATION

Corresponding Author

Meiqi Xu − *Key Laboratory of Novel Targets and Drug Study for Neural Repair of Zhejiang Province, School of Medicine, Hangzhou City University, Hangzhou 310015 Zhejiang, China;* orcid.org/0009-0000-2365-446X; Phone: +86-571-88284325; Email: xumq@hzcu.edu.cn; Fax: +86-571-88284325

Authors

Jiefu Pang − *School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018 Zhejiang, China*

Yangyang Ye − *Key Laboratory of Novel Targets and Drug Study for Neural Repair of Zhejiang Province, School of Medicine, Hangzhou City University, Hangzhou 310015 Zhejiang, China*

Ziyi Zhang − *Key Laboratory of Novel Targets and Drug Study for Neural Repair of Zhejiang Province, School of Medicine, Hangzhou City University, Hangzhou 310015 Zhejiang, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c01374

Author Contributions

M.X. conceived the basic idea and designed the framework. M.Q. and J.P. performed the experiments. M.X., J.P., and Z.Z. wrote the manuscript. Y.Y. revised the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Authors gratefully acknowledge the financial support from the Zhejiang City University Research and Incubation Fund (J-202326).

REFERENCES

(1) Siegel, R. L.; Miller, K. D.; Wagle, N. S.; et al. Cancer statistics, 2023. *CA Cancer J. Clin* **2023**, *73* (1), 17−48.

(2) Sung, H.; Ferlay, J.; Siegel, R. L.; et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin* **2021**, *71* (3), 209−249.

(3) Cui, C.; Merritt, R.; Fu, L.; et al. Targeting calcium signaling in cancer therapy. *Acta Pharm. Sin B* **2017**, *7* (1), 3−17.

(4) Shewach, D. S.; Kuchta, R. D. Introduction to cancer chemotherapeutics. *Chem. Rev.* **2009**, *109* (7), 2859−2861.

(5) Eghtedari, M.; Jafari Porzani, S.; Nowruzi, B. Anticancer potential of natural peptides from terrestrial and marine environments: A review. *Phytochemistry Letters* **2021**, *42*, 87−103.

(6) Qiao, X.; Wang, Y.; Yu, H. Progress in the mechanisms of anticancer peptides. *Shengwu Gongcheng Xuebao* **2019**, *35* (8), 1391−1400.

(7) Luan, X.; Wu, Y.; Shen, Y. W.; et al. Cytotoxic and antitumor peptides as novel chemotherapeutics. *Nat. Prod. Rep.* **2021**, *38* (1), 7−17.

(8) Divyashree, M.; Mani, M. K.; Reddy, D.; et al. Clinical Applications of Antimicrobial Peptides (AMPs): Where do we Stand Now? *Protein Pept. Lett.* **2020**, *27* (2), 120−134.

(9) Tornesello, A. L.; Borrelli, A.; Buonaguro, L.; et al. Antimicrobial Peptides as Anticancer Agents: Functional Properties and Biological Activities. *Molecules* **2020**, *25* (12), 2850.

(10) Leite, M. L.; da Cunha, N. B.; Costa, F. F. Antimicrobial peptides, nanotechnology, and natural metabolites as novel approaches for cancer treatment. *Pharmacol. Ther.* **2018**, *183*, 160−176.

(11) Yao, L.; Li, W.; Zhang, Y.; et al. Accelerating the Discovery of Anticancer Peptides through Deep Forest Architecture with Deep Graphical Representation. *Int. J. Mol. Sci.* **2023**, *24* (5), 4328.

(12) Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; et al. ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides. *Molecules* **2019**, *24* (10), 1973.

(13) Li, N.; Ainsworth, R.; Wu, M.; et al. MIEC-SVM: automated pipeline for protein peptide/ligand interaction prediction. *Bioinformatics (Oxford, England)* **2016**, *32* (6), 940−942.

(14) 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* (4), 1276−1314.

(15) Feng, H.; Wang, F.; Li, N.; et al. A Random Forest Model for Peptide Classification Based on Virtual Docking Data. *Int. J. Mol. Sci.* **2023**, *24* (14), 11409.

(16) Yang, C.; Ren, J.; Li, B.; et al. Identification of gene biomarkers in patients with postmenopausal osteoporosis. *Mol. Med. Rep.* **2019**, *19* (2), 1065−1073.

(17) Wang, H.; Zhao, J.; Zhao, H.; et al. CL-ACP: a parallel combination of CNN and LSTM anticancer peptide recognition model. *BMC Bioinf.* **2021**, *22* (1), 512.

(18) Yi, H.; You, Z.; Zhou, X.; et al. ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation. *Molecular therapy. Nucleic acids* **2019**, *17*, 1−9.

(19) Yang, X.; Jin, J.; Wang, R.; et al. CACPP: A Contrastive Learning-Based Siamese Network to Identify Anticancer Peptides Based on Sequence Only. *J. Chem. Inf. Model.* **2023**.

(20) Sun, M.; Hu, H.; Pang, W.; et al. ACP-BC: A Model for Accurate Identification of Anticancer Peptides Based on Fusion Features of Bidirectional Long Short-Term Memory and Chemically Derived Information. *Int. J. Mol. Sci.* **2023**, *24* (20), 15447.

(21) Zhang, L.; Zhang, C.; Gao, R.; et al. Prediction of aptamer-protein interacting pairs using an ensemble classifier in combination with various protein sequence attributes. *BMC Bioinf.* **2016**, *17* (1), 225.

(22) Zhong, T.; Yao, X.; Zhang, S.; et al. A self-assembling nanomedicine of conjugated linoleic acid-paclitaxel conjugate (CLA-PTX) with higher drug loading and carrier-free characteristic. *Sci. Rep.* **2016**, *6*, 36614.

(23) Xu, M.; Hao, Y.; Wang, J.; et al. Antitumor Activity of α-Linolenic Acid-Paclitaxel Conjugate Nanoparticles: In vitro and in vivo. *Int. J. Nanomed.* **2021**, *16*, 7269−7281.

(24) Zakharova, E.; Orsi, M.; Capecchi, A.; et al. Machine Learning Guided Discovery of Non-Hemolytic Membrane Disruptive Anticancer Peptides. *ChemMedChem* **2022**, *17* (17), No. e202200291.

(25) Alsanea, M.; Dukyil, A.; Afnan.; et al. To Assist Oncologists: An Efficient Machine Learning-Based Approach for Anti-Cancer Peptides Classification. *Sensors* **2022**, *22* (11), 4005.

(26) Charoenkwan, P.; Chiangjong, W.; Lee, V. S.; et al. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. *Sci. Rep.* **2021**, *11* (1), 3017.

(27) Wei, L.; Zhou, C.; Chen, H.; et al. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **2018**, *34* (23), 4007−4016.

(28) Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *EXCLI J* **2018**, *17*, 734−752.

https://doi.org/10.1021/acsomega.4c01374
ACS Omega 2024, 9, 16820−16831