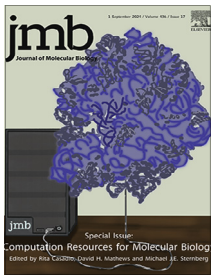

mACPpred 2.0: Stacked Deep Learning for Anticancer Peptide Prediction with Integrated Spatial and Probabilistic Feature Representations

Vinoth Kumar Sangaraju [1,†], Nhat Truong Pham [1,†], Leyi Wei [2], Xue Yu [3,*] and Balachandran Manavalan [1,*]

*1 - **Department of Integrative Biotechnology,** College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon, 16419, Gyeonggi-do, Republic of Korea*
*2 - **Faculty of Applied Sciences,** Macao Polytechnic University, Macau*
*3 - **Beidahuang Industry Group General Hospital,** 150001 Harbin, China*

***Correspondence to Xue YuBalachandran Manavalan:** yuxuexy@yeah.net (X. Yu), bala2022@skku.edu (B. Manavalan)*
https://doi.org/10.1016/j.jmb.2024.168687
Edited by Michael Sternberg

Abstract

Anticancer peptides (ACPs), naturally occurring molecules with remarkable potential to target and kill cancer cells. However, identifying ACPs based solely from their primary amino acid sequences remains a major hurdle in immunoinformatics. In the past, several web-based machine learning (ML) tools have been proposed to assist researchers in identifying potential ACPs for further testing. Notably, our meta-approach method, mACPpred, introduced in 2019, has significantly advanced the field of ACP research. Given the exponential growth in the number of characterized ACPs, there is now a pressing need to create an updated version of mACPpred. To develop mACPpred 2.0, we constructed an up-to-date benchmarking dataset by integrating all publicly available ACP datasets. We employed a large-scale of feature descriptors, encompassing both conventional feature descriptors and advanced pre-trained natural language processing (NLP)-based embeddings. We evaluated their ability to discriminate between ACPs and non-ACPs using eleven different classifiers. Subsequently, we employed a stacked deep learning (SDL) approach, incorporating 1D convolutional neural network (1D CNN) blocks and hybrid features. These features included the top seven performing NLP-based features and 90 probabilistic features, allowing us to identify hidden patterns within these diverse features and improve the accuracy of our ACP prediction model. This is the first study to integrate spatial and probabilistic feature representations for predicting ACPs. Rigorous cross-validation and independent tests conclusively demonstrated that mACPpred 2.0 not only surpassed its predecessor (mACPpred) but also outperformed the existing state-of-the-art predictors, highlighting the importance of advanced feature representation capabilities attained through SDL. To facilitate widespread use and accessibility, we have developed a user-friendly for mACPpred 2.0, available at https://balalab-skku.org/mACPpred2/.

Introduction

Cancer, a leading cause of death globally, is characterized by uncontrolled cell growth and spread. This complex disease disrupts the normal cellular life cycle, often resulting in malignant tumors and the potential for metastasis.[1] Given the complexity and diversity of cancer, developing effective treatments poses significant challenges, as each cancer type often requires a specific treat-

ment strategy.[2] In 2018, the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC) reported 18.1 million new cancer cases and ~9.6 million cancer-related deaths.[3] While conventional cancer treatments like radiotherapy and chemotherapy remain essential,[4] they have significant limitations. These methods are costly, often cause severe side effects on healthy cells, and face challenges with cancer cells developing resistance to chemotherapeutic drugs. These challenges highlight the critical need for novel and effective cancer therapies. In this context, peptide-based treatments are gaining attention as a promising alternative. They offer increased specificity, better tumor penetration, and lower toxicity, marking a new frontier in cancer treatment efforts.[5]

Anticancer peptides (ACPs), a sub-class of antimicrobial peptides (AMPs), are short chains of amino acids, typically 5–50 amino acids in length, and predominantly cationic. This composition gives them a distinctive ability to target and potentially eliminate cancer cells. Identifying potential ACPs through computational methods, before embarking on costly and time-consuming laboratory experiments, is a crucial step in accelerating drug discovery. Over the past decade, numerous ACP datasets have been constructed to aid in the development of ACP identification tools. These includes CpACpP,[6] mACPpred,[7] MLACP 2.0,[8] AntiCP 2.0 (Main) and (Alternate),[9] ACP-Mixed-80,[10] ACP-DL (740) and (240),[11] ACPred,[12] ACPred-FL,[13] and the LEE dataset.[14] Specifically, mACPpred was developed using a redundancy-reduced dataset and a meta-approach. This approach involved a two-step feature selection protocol applied to seven different feature encodings, resulting in optimal feature-based models for each encoding. The predicted probabilities of ACPs from these models were then used as feature vectors and input into a support vector machine for the final prediction. mACPpred has gained widespread popularity and use within the research community due to its effectiveness. However, with the exponential growth of experimentally verified ACPs, it is crucial to update mACPpred using advanced computational techniques to further improve its accuracy and robustness.

In this study, we introduced mACPpred 2.0, a significantly enhanced version of mACPpred, trained on the updated redundancy reduced larger training dataset. Briefly, peptide sequence converted into numerical feature vectors using 16 pre-trained natural language processing (NLP)-based embeddings and 15 conventional feature descriptors. These features were assessed using eleven ML-based classifiers, identifying the top seven NLP-based embeddings with the highest discriminative ability. Additionally, we observed that 184 out of 341 baseline models exceeded 80% accuracy. To construct the final predictor, we employed a stacking deep learning (SDL) framework with multiple 1D convolutional neural network (1D CNN) blocks. Each block utilized high-spatial feature representations derived from the top seven NLP-based embeddings. The fully connected (FC) layers from these blocks were then fused together and subsequently combined with the 90 probabilistic features from the best-performing baseline models. These features were processed through additional FC layers for the final classification. Rigorous cross-validation and independent tests demonstrated that mACPpred 2.0 outperformed the state-of-the-art predictors, including its predecessor, suggesting that the integration of SDL framework and NLP-based embeddings with probabilistic features from the optimal baseline models significantly enhances the effectiveness and robustness of mACPpred 2.0. We anticipate that the publicly available mACPpred 2.0 will be a valuable tool for researchers, enabling accurate ACP identification and streamlining cancer therapy and personalized drug development.

Materials and Methods

Benchmark dataset

This study aims to develop a prediction model using a dataset that overlaps with the training samples and to evaluate the proposed model using an independent dataset that does not overlap with the training samples of existing methods. Specifically, we compiled the training and independent samples from 11 existing methods, including ACP-DL, ACP-MLC,[15] ACPred-Fuse,[16] ACPred, AntiCP 2.0, CpACpP, MLACP, MLACP 2.0, mACPpred, ACPred-FL, and AMPfun,[17] and categorized into ACPs and non-ACPs. Notably, these methods utilized a unique dataset, which has been widely adopted as the training samples by existing ACP methods. Sequences containing non-standard amino acids (B, O, U, X, J, Z) were excluded, resulting in 7,261 ACPs and 13,023 non-ACPs. After removing identical sequences and those shorter than 5 or longer than 50 amino acids, the dataset was refined to 7,150 ACPs and 12,252 non-ACPs. It should be noted that high sequence-similarity in training samples can lead to over estimating model performance, a common issue in previously proposed ACP methods. To reduce sequence similarity, we applied CD-HIT[18] with a threshold of 0.85 on both training and independent samples, resulting in 1,176 ACPs and 4,001 non-ACPs. To prevent class imbalance during model training, we randomly selected 1,176 non-ACPs for the training set. Notably, the resulting training dataset is approximately 5 times larger than the previous version. The independent dataset, used for unbiased model evaluation, consists of 610 ACPs and 2,760 non-ACPs.

Feature extraction

To represent the peptide sequences into numerical vectors, we employed 16 pre-trained NLP-based embeddings and 15 conventional feature descriptors. The NLP-based embeddings have been trained on extensive protein sequence databases, enabling them to effectively capture feature representations from both peptide and protein sequences. They included Bepler, CPCProt, ESM, ESM1b (ESB), ESM1v (ESV), FastText, GloVe, PLUSRNN (PRNN), ProtTransAlbertBFD (PTAB), ProtTransBertBFD (PTBB), ProtTransT5BFD (PTB), ProtTransT5UniRef50 (PTU), ProtTransT5XLU50 (PTXU), ProtTransXLNetUniRef100 (PTXLU), Seq2Vec (S2V), and Word2Vec. The conventional feature descriptors encompassed amino acid composition (AAC), amphiphilic pseudo-amino acid composition (APAAC), peptide descriptor encoding (PDE), composition of k-spaced amino acid group pairs (CKSAAGP), composition-transition-distribution-composition (CTDC), composition-transition-distribution-distribution (CTDD), composition-transition-distribution-transition (CTDT), conjoint triad (CTriad), dipeptide deviation from expected mean (DDE), di-peptide composition (DPC), grouped dipeptide composition (GDPC), grouped tripeptide composition (GTPC), pseudo-amino acid composition (PAAC), quasi-sequence-order (QSOrder), and tri-peptide-composition (TPC). Comprehensive details about these embeddings and descriptors as well as their processing can be found in the supplementary information.

Model development

In mACPpred 2.0, we designed an SDL framework by integrating multiple 1D CNN blocks for the identification of ACPs (Figure 1). This structure is specifically aimed at learning and extract high-spatial information from the top seven NLP-based embeddings. Each 1D CNN block comprised of three 1D convolution (Conv1D) layers, each followed by three rectified linear unit (ReLU) activation functions. Notably, when given an input feature X with E elements, $X = \{x_1, x_2, x_3, ..., x_E\}$, along with the number of filters F, kernel size K defined as $K = \{k_1, k_2, k_3, ..., k_F\}$, and bias b, the output feature map FM with Z elements after convolution, denoted as $FM = \{fm_1, fm_2, fm_3, ..., fm_Z\}$, each of which can be computed as follows:

$$fm_i = \sum_{f=0}^{F-1} k_f \cdot x_{i+f} + b, \quad (1)$$

where i in range [1, Z]. After that, every feature map undergoes a ReLU activation function to perform element-wise non-linear calculation, as follows:

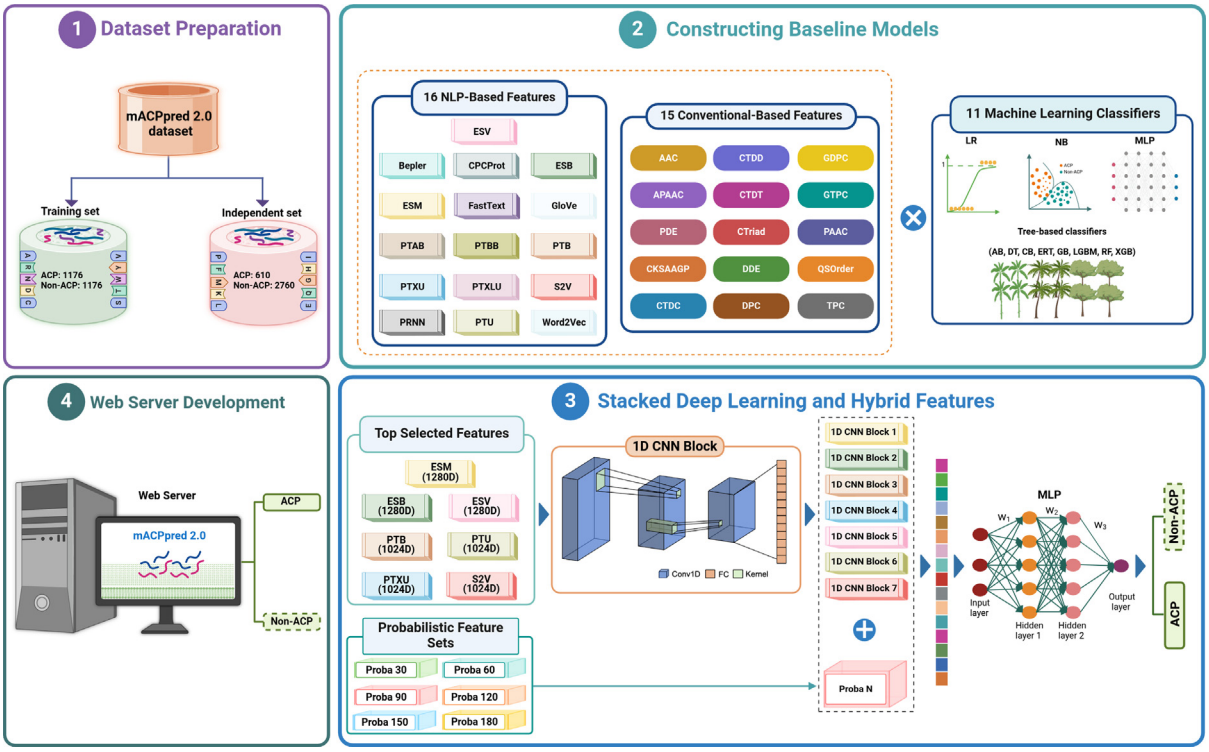

Figure 1. Overview of the four main phases in the development of mACPpred 2.0: (i) dataset preparation, (ii) feature extraction and baseline model construction, (iii) stacked deep learning with hybrid features for enhanced prediction accuracy, and (iv) web server development.

$$a_i = ReLU(fm_i), \quad (2)$$

where a_i represents the output of the ReLU activation function. The operation of the ReLU activation function is defined as follows:

$$ReLU(fm_i) = \max(fm_i, 0). \quad (3)$$

Importantly, the kernel sizes for the first, second, and third Conv1D layers were selected as 5, 3, and 3, respectively. The resulting output feature maps from these layers are then flattened and fed into a FC layer, representing the feature representation of each NLP-based embedding input. After stacking seven 1D CNN blocks, the final FC layers of these blocks were fused together by concatenation. These high-spatial feature representations were subsequently integrated with the top 90 probabilistic features derived from the selected top single-feature baseline models to construct hybrid features. These hybrid features were then fed into additional FC layers for final classification. The SDL framework in mACPpred 2.0 containing various DL parameters that were fine-tuned through a comprehensive hyperparameter search and assessed their performance using five-fold cross-validation. More specifically, mACPpred 2.0 was implemented using TensorFlow and Scikit-learn. The specific search ranges utilized to obtain the optimal models for mACPpred 2.0 are detailed in Table S1.

Performance evaluation metrics

To thoroughly evaluate our proposed method's effectiveness and compare with the state-of-the-art predictors, we employed seven widely used performance metrics,[19,20] including Matthews correlation coefficient (MCC), sensitivity (Sn), specificity (Sp), precision (PRE), ACC, area under the receiver operating characteristic curve (AUC), and F1-score. The mathematical expressions for these metrics are defined as follows:

$$MCC = \frac{N_{TP} \times N_{TN} - N_{FP} \times N_{FN}}{\sqrt{(N_{TP} + N_{FP}) \times (N_{TP} + N_{FN}) \times (N_{TN} + N_{FP}) \times (N_{TN} + N_{FN})}}, \quad (4)$$

$$Sn = \frac{N_{TP}}{N_{TP} + N_{FN}}, \quad (5)$$

$$Sp = \frac{N_{TN}}{N_{TN} + N_{FP}}, \quad (6)$$

$$PRE = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad (7)$$

$$ACC = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}}, \quad (8)$$

$$F1 - score = \frac{2 \times N_{TP}}{2 \times N_{TP} + N_{FP} + N_{FN}}, \quad (9)$$

where N_{TP}, N_{FP}, N_{TN}, and N_{FN} denote the number of true-positive, false-positive, true-negative, and false-negative samples, respectively.

Results and Discussion

Comparative analysis of NLP-based embeddings and conventional feature descriptors in ACP prediction using 11 conventional ML algorithms

We conducted a large-scale analysis of feature descriptors, including 16 NLP-based embeddings and 15 conventional feature descriptors (see Methods section). To assess their effectiveness in ACP prediction, we employed 11 different ML algorithms, generating a total of 341 single-feature baseline models (Table S2). Notably, the hyperparameters of all classifiers were optimized using five-fold cross validation. Figures 2A and B respectively show the cross-validation performance (in terms of ACC) of 341 single-feature baseline models based on 16 NLP-based embeddings and 15 conventional feature descriptors. Interestingly, out of the 341 baseline models, 184 models achieved an ACC exceeding 80%. However, we observed that performance varied significantly across classifiers and features. For instance, models based on PTU exhibited ACC range, from 0.728 to 0.840. We considered all these top performing models for further analysis. To assess each feature's discriminative potential, we calculated the average performance (in terms of MCC) of the eleven classifier-based models (Figure 2C). The results show that seven NLP-based embeddings demonstrated significantly higher discriminative ability, with MCC values exceeding 0.625, compared to the others. Consequently, we selected these top performing embeddings, including ESB, ESV, ESM, PTB, PTU, PTXU, and S2V, for further analysis. Inspired by previous studies that used probabilistic features to construct meta-learning and adaptive feature representation learning frameworks,[8,21,22] we also extracted probabilistic features from the 184 single-feature baseline models. Integrating these seven NLP-based embeddings and probabilistic features to construct the SDL framework.

Development of mACPpred 2.0 using SDL framework with hybrid probabilistic features

Instead of linearly combining the seven NLP-based embeddings, we leveraged the high-spatial feature representations learned from each embedding using 1D CNN block. This approach allowed us to capture complex patterns and relationships within the embeddings. The feature representations generated by these seven 1D CNN blocks were then integrated with the probabilistic features derived from the high-performing single-feature models. It is unclear whether all baseline models or only a subset are necessary for optimal performance. To investigate this, we ranked the probabilistic features by the

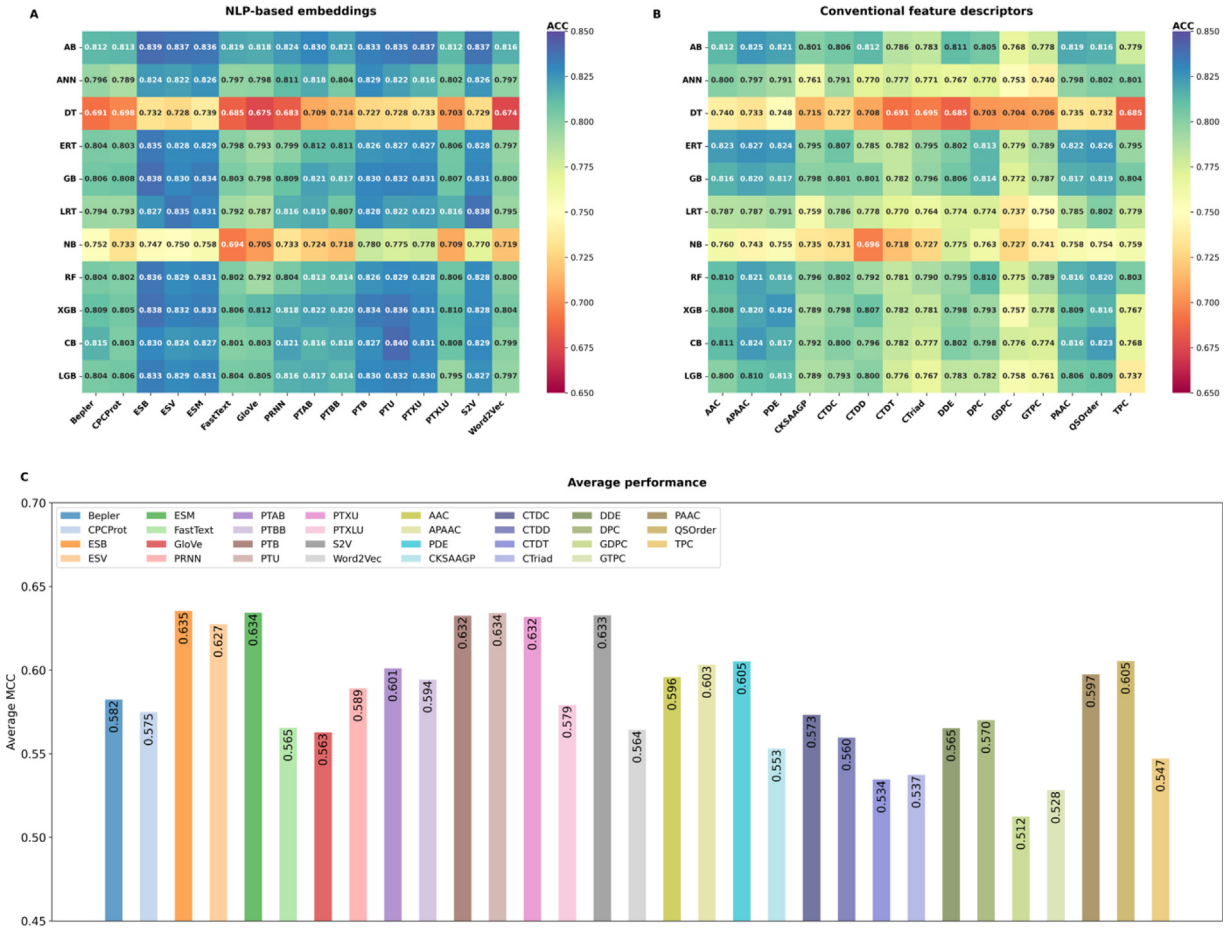

Figure 2. Performance comparison in terms of accuracy (ACC) for eleven conventional ML-based classifiers using 16 NLP-based embeddings **(A)** and 15 conventional feature descriptors **(B)**. Average Matthews correlation coefficient (MCC) for each feature **(C)**.

ACC and generated six subsets ranging from 30 to 180 probabilistic features, with an interval of 30.

Figure 3A depicts the cross-validation performance of the top seven NLP-based embeddings, combined with different subsets of probabilistic features. The integrating probabilistic features with spatial feature representation learned from these seven 1D CNN blocks significantly improved the prediction performance compared to the spatial features obtained from NLP-based embeddings. Notably, models using the top seven NLP-based embeddings with the top 90 and 150 probabilistic features achieved the highest MCC values of 0.810 during training. However, to minimize the computational time for analyzing any unseen peptides, we selected a model (mACPpred 2.0) that combines the spatial feature representation with the top 90 probabilistic features. The selected 90 probabilistic features predominantly originated from NLP-based embeddings accounting for 74 models (82.2%), while conventional feature descriptor contributed 16 models (17.78%). The classifiers contributing these models were diverse, including CB (11 models), AB (14 models), XGB (12 models), LRT (8 models), ERT (12 models), LGB (7 models), RF (9 models), ANN (7 models), and GB (10 models). Notably, mACPpred 2.0 obtained Sn, Sp, ACC, and AUC values of 0.912, 0.898, 0.905, and 0.965, respectively. Specifically, mACPpred 2.0 performed better than the other models in Figure 3A, with improvements ranging from 1.10 to 12.90% in MCC, 0.10 to 6.40% in ACC, and 0.20 to 5.90% in AUC. To show the advantage of SDL framework, we compared mACPpred 2.0 with the best single-feature baseline model (CB with PTU). The SDL approach resulted in remarkable improvements for mACPpred 2.0, including 12.90% increase in MCC, 8.0% in Sn, 5.0% in Sp, 6.50% in ACC, and 5.60% in AUC.

To ensure the robustness of our trained models (Figure 3A), it is crucial to evaluate its performance on an independent imbalanced dataset, as this reflects real-world scenarios more accurately. Figure 3B shows that models using hybrid features (NLP-based embeddings and their combination with different probabilistic features) maintained performance levels similar to those

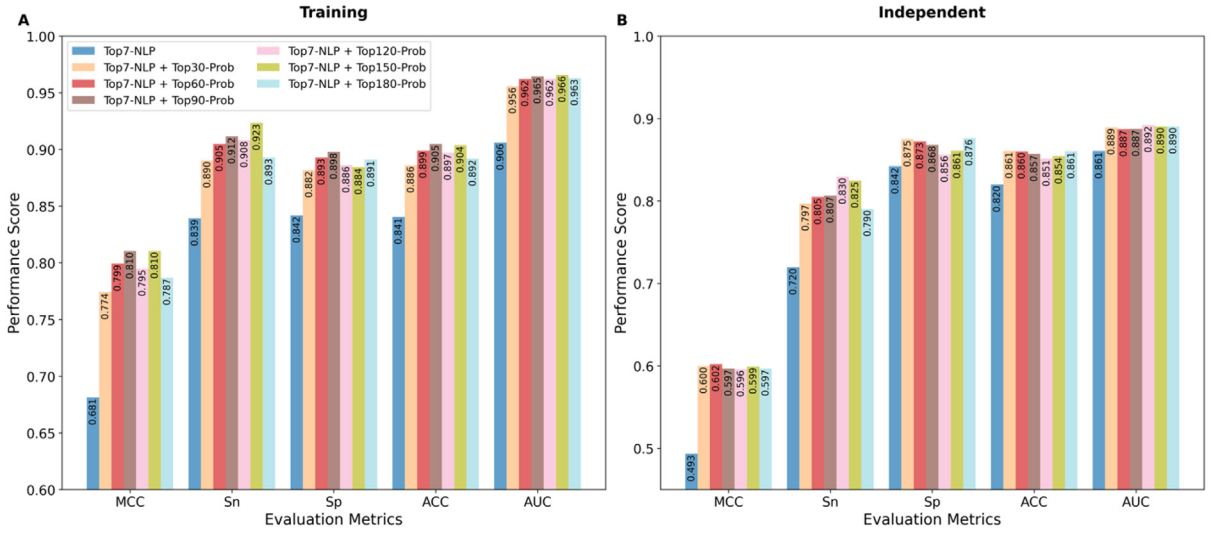

Figure 3. Performance comparison of the top seven NLP-based embeddings with and without different probabilistic features on training **(A)** and independent **(B)** datasets.

observed during cross-validation. Notably, our final model, mACPpred 2.0 achieved MCC, Sn, Sp, ACC, and AUC values of 0.597, 0.807, 0.868, 0.857, and 0.887, respectively. Compared to the model based solely on spatial feature representation (NLP-based embeddings), mACPpred 2.0 achieved significant improvements with 10.40% in MCC, 8.70% in Sn, 2.60% in Sp, 3.70% in ACC, and 2.60% in AUC. Overall, mACPpred 2.0 consistently outperformed other models on both training and independent datasets, demonstrating superior performance compared to models relying solely on spatial feature representations. mACPpred 2.0 success can be attributed to the effectiveness of the SDL framework, which integrates the spatial contextual information from NLP-based embeddings and the meta-information extracted from probabilistic features of baseline models.

Performance comparison between mACPpred 2.0 and publicly available ACP predictors on the independent dataset

We comprehensively evaluated mACPpred 2.0 and its predecessor (mACPpred) as well as 13 publicly available ACP predictors on the independent dataset. These include MLACP 2.0,[8] iAMPCN,[23] ACP-MHCNN[24] (two models), ACPred-BMF[25] (two models), ACPred,[12] AMPfun,[17] iAMP-RAAC,[26] AntiCP 2.0[9] (two models), CancerGram,[27] PreTP-2L,[28] iDACP[29] (five models), iACP-DRLF,[30] TriNet,[31] and mACPpred.[7] Table 1 shows that mACPpred 2.0 achieved the best performance across all global metrics, with MCC, ACC, AUC, and F1-score values of 0.597, 0.857, 0.887, and 0.671, respectively. In comparison, its predecessor obtained MCC, Sn, Sp, PRE, ACC, AUC, and F1-score values of 0.489, 0.754, 0.819, 0.479, 0.807, 0.842, and 0.586, all of which are consistently lower than those of mACPpred 2.0. Notably, even though iAMPCN was recently introduced and showed better performance than most existing ACP predictors, it has the lowest ACC on our independent dataset. Overall, mACPpred 2.0 outperformed its predecessor and all existing predictors on the independent dataset, highlighting the effectiveness of our proposed method.

Analyzing feature contribution with UMAP visualization

To gain deeper insights into the learned feature representations that contributed to the improved performance of mACPpred 2.0 in discriminating ACPs from non-ACPs, we employed UMAP[32] to visualize the extracted feature representations from each component in the mACPpred 2.0 model on both the training and independent datasets (Figure S1). The visualizations reveal that the feature representations learned by 1D CNN blocks with the seven NLP-based embeddings exhibit significant overlap between ACPs and non-ACPs. However, when these feature representations are combined with the top 90 probabilistic features, mACPpred 2.0 demonstrates a markedly improved discriminative ability, as evidenced by the clearer separation between ACP and non-ACP samples. Interestingly, a similar patterns were observed on the independent dataset. This demonstrates that the model effectively learns robust patterns that can significantly discriminate ACPs from non-ACPs in the training dataset, and these patterns are transferable and generalizable to new independent dataset.

Table 1 Performance comparison of mACPpred 2.0, its predecessor (mACPpred), and other publicly available predictors on an independent dataset.

Method	MCC	Sn	Sp	PRE	ACC	AUC	F1-score
mACPpred 2.0	0.597	0.807	0.868	0.575	0.857	0.887	0.671
iAMPCN	−0.213	0.625	0.156	0.141	0.241	0.681	0.230
mACPpred	0.489	0.754	0.819	0.479	0.807	0.842	0.586
MLACP 2.0	0.557	0.792	0.849	0.536	0.838	0.893	0.639
ACP-MHCNN (740)	0.132	0.357	0.789	0.272	0.711	–	0.309
ACP-MHCNN (500/164)	−0.102	0.167	0.716	0.115	0.616	–	0.136
ACPred-BMF (Main)	0.198	0.672	0.584	0.263	0.600	0.326	0.378
ACPred-BMF (Alternate)	0.251	0.441	0.829	0.364	0.759	0.263	0.399
ACPred	0.278	0.510	0.804	0.365	0.751	0.739	0.426
AMPfun	0.429	0.769	0.758	0.412	0.760	0.802	0.537
iAMP-RAAC	0.102	0.066	0.979	0.408	0.814	–	0.113
AntiCP 2.0 (Main)	0.178	0.433	0.772	0.295	0.710	0.597	0.351
AntiCP 2.0 (Alternate)	0.472	0.605	0.887	0.543	0.836	0.835	0.572
CancerGram	0.110	0.066	0.981	0.435	0.815	0.673	0.114
iDACP (Sp of 90%)	0.255	0.203	0.963	0.546	0.825	0.852	0.296
iDACP (Sp of 80%)	0.330	0.356	0.927	0.519	0.824	0.852	0.422
iDACP (Sp of 70%)	0.380	0.484	0.892	0.497	0.818	0.852	0.490
iDACP (Sp of 60%)	0.440	0.616	0.861	0.495	0.817	0.852	0.549
iDACP (Sp of 50%)	0.451	0.680	0.833	0.474	0.805	0.852	0.559
TriNet	0.229	0.433	0.817	0.343	0.747	0.718	0.383
PreTP-2L	0.065	0.021	0.994	0.448	0.818	–	0.041

Web server development

To ensure widespread accessibility and usability of mACPpred 2.0, we have deployed a user-friendly web server at https://balalab-skku.org/mACPpred2/. Additionally, to promote reproducibility and enable comparisons, we have provided all newly constructed training and imbalanced independent datasets used in our study. To help users navigate mACPpred 2.0 for prediction, the help page is available at https://balalab-skku.org/mACPpred2/help/, guiding users through the steps of data formatting, submission, result interpretation, and retrieving submitted jobs. It is important for users to format peptide sequences in FASTA format prior to submitting the server, and we provide example sequences to aid in this process. Once a submission is made, the results are displayed in table format, and users can download these results as CSV files for further analysis. Moreover, we have provided the standalone program for mACPpred 2.0 at https://github.com/nhattruongpham/mACPpred2/ for local deployment.

Conclusion

This study introduces mACPpred 2.0, a significantly improved predictor for identifying ACPs. The enhanced performance is primarily attributed due to the SDL approach, which leverages both the spatial context captured from NLP-based embeddings and the meta-information extracted from probabilistic features of baseline models. Briefly, SDL extracts high-spatial feature representations from NLP-based embeddings using multiple 1D CNN blocks. These extracted feature representations were then combined and integrated with the probabilistic features extracted from the top single-feature baseline models to build the final predictive model. Rigorous cross-validation and independent testing demonstrated mACPpred 2.0 consistent performance, underscoring its generalizability and effectiveness. Additionally, mACPpred 2.0 significantly outperformed existing predictors, making it the

best available predictor at the moment. It is freely accessible as both standalone program and web server. Notably, the SDL approach can be applied to other sequence-based function prediction problems,[33,34] including genomic DNA function prediction (e.g., enhancers and replication origins), and other peptide therapeutic function predictions. In future studies, we plan to expand our dataset by collecting more non-ACPs, considering various scenarios as employed in MLACP 2.0. We also aim to implement novel feature descriptors based on positional information and explore different DL-based frameworks, including contrastive-learning, multi-task learning, meta-learning methods, stacking frameworks, and deep ensemble learning methods.

Funding

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2021R1A2C1014338, RS-2023-00217881, RS-2024-00344752). This research was supported by the Department of Integrative Biotechnology, Sungkyunkwan University (SKKU) and the BK21 FOUR Project. The work is supported by Internal Research Grants of Macao Polytechnic University (no. RP/CAI02/2023) and the Science and Technology Development Fund (no. 0177/2023/RIA3).

CRediT authorship contribution statement

Vinoth Kumar Sangaraju: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation. **Nhat Truong Pham:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Leyi Wei:** Writing – review & editing, Formal analysis, Conceptualization. **Xue Yu:** Supervision, Validation, Writing – review & editing. **Balachandran Manavalan:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Funding acquisition, Conceptualization.

DATA AVAILABILITY

The web server for mACPpred 2.0, along with all the newly constructed training and independent datasets, is publicly accessible at https://balalab-skku.org/mACPpred2/. Additionally, the standalone program is available at https://github.com/nhattruongpham/mACPpred2/ for local deployment. Furthermore, we have provided the baseline and final models via Zenodo at https://doi.org/10.5281/zenodo.11350064.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Korea Bio Data Station (K-BDS) with computing resources including technical support. We thank Nattanong Bupi for his assistance in creating the figure, and CBBL members for their valuable discussion.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.jmb.2024.168687.

Received 8 February 2024;
Accepted 20 June 2024;
Available online 25 June 2024

Keywords:
anticancer peptides;
stacking deep learning;
pre-trained natural language processing-based embeddings

† Equally contributed to this work.

References

1. Anand, U., Dey, A., Chandel, A.K.S., Sanyal, R., Mishra, A., Pandey, D.K., et al., (2023). Cancer chemotherapy and beyond: Current status, drug candidates, associated risks and progress in targeted therapeutics. *Genes Dis.* **10**, 1367–1401.
2. Basith, S., Cui, M., Macalino, S.J., Choi, S., (2017). Expediting the design, discovery and development of anticancer drugs using computational approaches. *Curr. Med. Chem.* **24**, 4753–4778.
3. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424.
4. Abbas, Z., Rehman, S., (2018). An overview of cancer treatment modalities. *Neoplasm.* **1**, 139–157.
5. Harris, F., Dennison, S.R., Singh, J., Phoenix, D.A., (2013). On the selectivity and efficacy of defense peptides with respect to cancer cells. *Med. Res. Rev.* **33**, 190–234.
6. Nasiri, F., Atanaki, F.F., Behrouzi, S., Kavousi, K., Bagheri, M., (2021). CpACpP: in silico cell-penetrating anticancer peptide prediction using a novel bioinformatics framework. *ACS Omega* **6**, 19846–19859.
7. Boopathi, V., Subramaniyam, S., Malik, A., Lee, G., Manavalan, B., Yang, D.C., (2019). mACPpred: A support

vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* **20**
8. Phan, L., Park, H.W., Pitti, T., Madhavan, T., Jeon, Y.J., Manavalan, B., (2022). MLACP 2.0: An updated machine learning tool for anticancer peptide prediction. *Comput Struct Biotec.* **20**, 4473–4480.
9. Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N., Raghava, G.P.S., (2021). AntiCP 2.0: An updated model for predicting anticancer peptides. *Brief. Bioinform.* **22**
10. He, W., Wang, Y., Cui, L., Su, R., Wei, L., (2021). Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides. *Bioinformatics* **37**, 4684–4693.
11. Yi, H.C., You, Z.H., Zhou, X., Cheng, L., Li, X., Jiang, T.H., et al., (2019). ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Mol. Ther.-Nucl. Acids* **17**, 1–9.
12. Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V., Shoombuatong, W., (2019). ACPred: A computational tool for the prediction and analysis of anticancer peptides. *Molecules* **24**
13. Wei, L.Y., Zhou, C., Chen, H.R., Song, J.N., Su, R., (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **34**, 4007–4016.
14. Manavalan, B., Basith, S., Shin, T.H., Choi, S., Kim, M.O., Lee, G., (2017). MLACP: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **8**, 77121–77136.
15. Deng, H., Ding, M., Wang, Y.M., Li, W.H., Liu, G.X., Tang, Y., (2023). ACP-MLC: A two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types. *Comput. Biol. Med.* **158**
16. Rao, B., Zhou, C., Zhang, G.Y., Su, R., Wei, L.Y., (2020). ACPred-Fuse: Fusing multi-view information improves the prediction of anticancer peptides. *Brief. Bioinform.* **21**, 1846–1855.
17. Chung, C.R., Kuo, T.R., Wu, L.C., Lee, T.Y., Horng, J.T., (2020). Characterization and identification of antimicrobial peptides with different functional activities. *Brief. Bioinform.* **21**, 1098–1114.
18. Fu, L.M., Niu, B.F., Zhu, Z.W., Wu, S.T., Li, W.Z., (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152.
19. Xie, X., Wu, C., Hao, Y., Wang, T., Yang, Y., Cai, P., et al., (2023). Benefits and risks of drug combination therapy for diabetes mellitus and its complications: a comprehensive review. *Front. Endocrinol. (Lausanne)* **14**, 1301093.
20. Zhu, W., Yuan, S.S., Li, J., Huang, C.B., Lin, H., Liao, B., (2023). A first computational frame for recognizing heparin-binding protein. *Diagnostics (Basel)* **13**
21. Pham, N.T., Phan, L., Seo, J., Kim, Y., Song, M.K.Y., Lee, S.K., et al., (2024). Advancing the accuracy of SARS-CoV-2 phosphorylation site detection via meta-learning approach. *Brief. Bioinform.* **25**
22. Pham, N.T., Terrance, A.T., Jeon, Y.-J., Rakkiyappan, R., Manavalan, B., (2024). ac4C-AFL: A high-precision identification of human mRNA N4-acetylcytidine sites based on adaptive feature representation learning. *Mol. Ther. Nucleic Acids* **35**
23. Xu, J., Li, F.Y., Li, C., Guo, X.D., Landersdorfer, C., Shen, H.H., et al., (2023). iAMPCN: A deep-learning approach for identifying antimicrobial peptides and their functional activities. *Brief. Bioinform.* **24**
24. Ahmed, S., Muhammod, R., Khan, Z.H., Adilina, S., Sharma, A., Shatabda, S., et al., (2021). ACP-MHCNN: An accurate multi-headed deep-convolutional neural network to predict anticancer peptides. *Sci. Rep.-Uk* **11**
25. Han, B.Q., Zhao, N., Zeng, C.S., Mu, Z.C., Gong, X.Q., (2022). ACPred-BMF: Bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction. *Sci. Rep.-Uk* **12**
26. Dong, G.F., Zheng, L., Huang, S.H., Gao, J., Zuo, Y.C., (2021). Amino acid reduction can help to improve the identification of antimicrobial peptides and their functional activities. *Front. Genet.* **12**
27. Burdukiewicz, M., Sidorczuk, K., Rafacz, D., Pietluch, F., Bakala, M., Slowik, J., et al., (2020). CancerGram: An effective classifier for differentiating anticancer from antimicrobial peptides. *Pharmaceutics* **12**
28. Yan, K., Guo, Y., Liu, B., (2023). PreTP-2L: Identification of therapeutic peptides and their types using two-layer ensemble learning framework. *Bioinformatics* **39**, btad125.
29. Huang, K.-Y., Tseng, Y.-J., Kao, H.-J., Chen, C.-H., Yang, H.-H., Weng, S.-L., (2021). Identification of subtypes of anticancer peptides based on sequential features and physicochemical properties. *Sci. Rep.-Uk* **11**, 13594.
30. Lv, Z., Cui, F., Zou, Q., Zhang, L., Xu, L., (2021). Anticancer peptides prediction with deep representation learning features. *Brief. Bioinform.* **22**, bbab008.
31. Zhou, W.Y., Liu, Y.F., Li, Y.X., Kong, S.Q., Wang, W.L., Ding, B.Y., et al., (2023). TriNet: A tri-fusion neural network for the prediction of anticancer and antimicrobial peptides. *Patterns* **4**
32. McInnes, L., Healy, J., Saul, N., Großberger, L., (2018). UMAP: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861.
33. Zou, X., Ren, L., Cai, P., Zhang, Y., Ding, H., Deng, K., et al., (2023). Accurately identifying hemagglutinin using sequence information and machine learning methods. *Front. Med. (Lausanne)* **10**, 1281880.
34. Zulfiqar, H., Guo, Z., Ahmad, R.M., Ahmed, Z., Cai, P., Chen, X., et al., (2023). Deep-STP: A deep learning-based approach to predict snake toxin proteins by using word embeddings. *Front. Med. (Lausanne)* **10**, 1291352.