Journal, 2024, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

BBATProt : A Framework Predicting Biological Function with Enhanced Feature Extraction via Explainable Deep Learning

Youqing Wang,[1] Xukai Ye,[1] Yue Feng,[2] Haoqian Wang,[1] Xiaofan Lin,[3,4] Xin Ma[1*] and Yifei Zhang[3,4*]

[1]College of Information Sciences and Technology, Beijing University of Chemical Technology, North Third Ring Road 15, 100029, Beijing, China, [2]College of Life Science and Technology, Beijing University of Chemical Technology, North Third Ring Road 15, 100029, Beijing, China, [3]State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, North Third Ring Road 15, 100029, Beijing, China and [4]Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, North Third Ring Road 15, 100029, Beijing, China.
*Corresponding author email: 2023700002@buct.edu.cn; yifeizhang@mail.buct.edu.cn
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Accurately predicting the functions of peptides and proteins from their amino acid sequences is essential for understanding life processes and advancing biomolecule engineering. Due to the time-consuming and resource-intensive nature of experimental procedures, computational approaches, especially those based on machine learning frameworks, have garnered significant interest. However, many existing machine learning tools are limited to specific tasks and lack adaptability across different predictions. Here we propose a versatile framework BBATProt for the prediction of various protein and peptide functions. BBATProt employs transfer learning with a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model, to effectively capture high-dimensional features from amino acid sequences. The whole custom-designed network, integrating Bidirectional Long Short-Term Memory (Bi-LSTM) and Temporal Convolutional Networks (TCN), can align with the spatial characteristics of proteins. It combines local and global feature extraction through attention mechanisms for precise functional prediction. This approach ensures that key features are adaptively extracted and balanced across diverse tasks. Comprehensive evaluations show BBATProt outperforms state-of-the-art models in predicting functions like hydrolytic catalysis, activity of peptides, and post-translational modification sites. Visualizations of feature evolution and refinement via attention mechanisms validate the framework's interpretability, providing transparency into the evolutional process and offering deeper insights into function prediction.

Introduction

Protein and peptide, the biomacromolecules composed of amino acid strings, execute a diverse array of essential biological functions in living organisms. Unraveling how these strings encode the structural and topological characteristics that determine their bioactivities is the core of biology. Previous studies, relying heavily on cumbersome wet experiments such as protein crystallography and biochemical assays, have advanced slowly (1). For instance, of the more than 100 million sequences recorded in the UniProt database, less approximately 0.5% have undergone manual annotation (UniProtKB/Swiss-Prot) (2). Despite AlphaFold's accomplishment in demonstrating that amino acid sequences encompass all necessary information for folding, accurately predicting their functions purely from these sequences remains difficult (3).

The difficulty in directly mapping a sequence of amino acids to its potential function arises from the complexity and diversity of functions (4). A particular biological function relies not only on the specific composition or configuration of key residues but is also influenced in varying degrees by other residues in the vicinity or even far away (5). For instance, the catalytic activity of an enzyme is mostly controlled by the residues in the catalytic pocket but may also be regulated by remote residues. The reactivity of a post-translational modification (PTM) site of a protein depends on the chemical environment of the target functional group, as well as the local landscape. Since peptides are considerably smaller than proteins, factors such as their composition, sequence patterns, and overall shapes all contribute to biological functions to different extents. Therefore, accurately predicting biological functions based on sequences requires an adaptive framework capable of automatically balancing multiple-level features of chemical and structural characteristics (6).

© The Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Traditional shallow machine learning methods, including K-Nearest Neighbors (KNN), Random Forest (RF), support vector machine (SVM) and so on (7; 8; 9; 10), have been commonly utilized for predicting specific functions of peptides and proteins. However, these methods suffer from limited generalization capabilities due to their dependence on high-quality manually annotated features, which restrict their ability to capture implicit and high-dimensional information within sequences, ultimately hindering their adaptability to diverse datasets and novel contexts. In contrast, deep neural network approaches have emerged to effectively map the complex, high-dimensional, and nonlinear relationships between sequences and their biological functions. These methods include Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Temporal Convolutional Networks (TCN), and advanced Bidirectional Encoder Representations from Transformers (BERT) (11; 12; 13; 14; 15). Notably, transfer learning techniques, such as BERT, which excel in leveraging contextual understanding from one domain to another, have proven highly effective in adapting to protein sequence analysis, enabling the model to capture intricate biological patterns with minimal task-specific data. Rives et al. used 250 million protein sequences to pretrain a protein language model by considering amino acid sequences as sentences (16), and demonstrated BERT's efficacy in capturing high-dimensional structural and functional information in protein sequences. Building on this architecture, models such as ProtBERT, ESM-1b, and BERT-Protein have been developed. These models extract high-dimensional abstract characteristics from protein sequences by using large-scale unsupervised learning and transfer learning methodologies (17; 18; 19).

Given the complexities inherent in protein and peptide function prediction, the development of a novel framework capable of extracting multi-level features that capture both local and global sequence information is imperative. In this study, we present BBATProt, a feature-enhanced network framework that utilizes BERT for effective dynamic word embedding extraction from amino acid sequences. The framework was custom-designed to incorporate a temporal data-based architecture that aligns with the interpretability of protein spatial structural relationships, thereby significantly enhancing the precision of function prediction. The workflow of BBATProt is illustrated in Fig. 1. To address the intricate interrelationships among high-dimensional features within sequences, BBATProt employs BERT for feature extraction, which adeptly captures contextual information inherent in the amino acid sequences. The incorporation of BERT facilitates the dynamic representation of sequences, allowing the model to adaptively learn from a vast array of protein sequences without the need for extensive prior knowledge of protein structures or functional domains. BBATProt's design emphasizes multi-level feature extraction, which is crucial for understanding the diverse functionalities of proteins and peptides. The framework utilizes a targeted combination of CNN, Bi-LSTM, and TCN is proposed to effectively encapsulate the complex information embedded in the encoded features. This methodology proficiently manages long-range dependencies and high-dimensional abstract features, optimizing the extraction of correlations among features derived from temporal data while minimizing computational overhead. Furthermore, the self-attention mechanism optimally leverages interdependencies among features, reflecting the nonlinear significance of different sequence positions for functional realization. To assess the algorithm's effectiveness, extensive evaluations across five independent datasets—including one hydrolytic enzyme dataset, two peptide datasets, and two post-translational modification (PTM) site datasets—reveal that BBATProt excels in accuracy, robustness, and generalization compared to current state-of-the-art (SOTA) models. Additioinaly to further validate the interpretability of the BBATProt architecture, t-distributed stochastic neighbor embedding (t-SNE) is used to illustrate the layer-by-layer evolution of features, while the refinement achieved through the attention mechanism is demonstrated separately. This study, grounded in dynamic word embeddings and neural networks, effectively identifies key functional features and showcases broad applicability and stability in practical scenarios, thus laying a solid foundation for future research in protein and peptide function prediction.

Table 1. Information on five peptide and protein datasets

Dataset	Training datasets		Independent datasets	
	Pos	Neg	Pos	Neg
Hydrolytic enzyme[25]	3132	3132	348	348
Antimicrobial peptide[26]	5536	5536	1536	1536
Inhibitory peptide[27]	532	532	133	133
Lysine crotonylation[28]	6975	6975	2989	2989
Lysine glycation[29]	3573	3573	200	200

Materials and Methods

Dataset Construction

Due to the lack of a benchmark dataset for validating the model's general versatility, Due to the lack of a benchmark dataset for validating the prediction model's general versatility, a comprehensive series of comparative experiments was carried out across multiple benchmark datasets to illustrate the effectiveness of BBATProt. This study seeks to conduct a comprehensive assessment of the algorithm's adaptability and robustness in a range of biological contexts. The framework was strategically designed to accommodate a range of functional prediction needs, encompassing instances such as antimicrobial peptides (AMP), dipeptidyl peptidase-IV (DPP-IV) inhibitory peptides, carboxylesterase hydrolysis prediction, and post-translational modification (PTM) site prediction (20; 21; 22; 23). To minimize redundancy within these datasets, a unique two-step operation deploying CD-HIT was executed at both protein and fragment levels (24). Table 1 presents the specific sources and detailed information of the datasets.

In the experiment, a 10-fold cross-validation approach was employed to evaluate the algorithm's robustness and reliability. This methodology entails partitioning the dataset into ten distinct, non-overlapping subsets. During each iteration, training is conducted using nine subsets, while the remaining subset is reserved for independent testing. This iterative process facilitates a thorough assessment of the algorithm's performance. The adoption of this cross-validation method helps minimize bias in the selection of training and test sets, thereby improving the scientific objectivity of the experimental results and ensuring replicability.

BERT feature extraction

Pre-trained natural language processing models have found extensive application across various fields (30; 31). Transfer learning techniques enable the execution of specific information

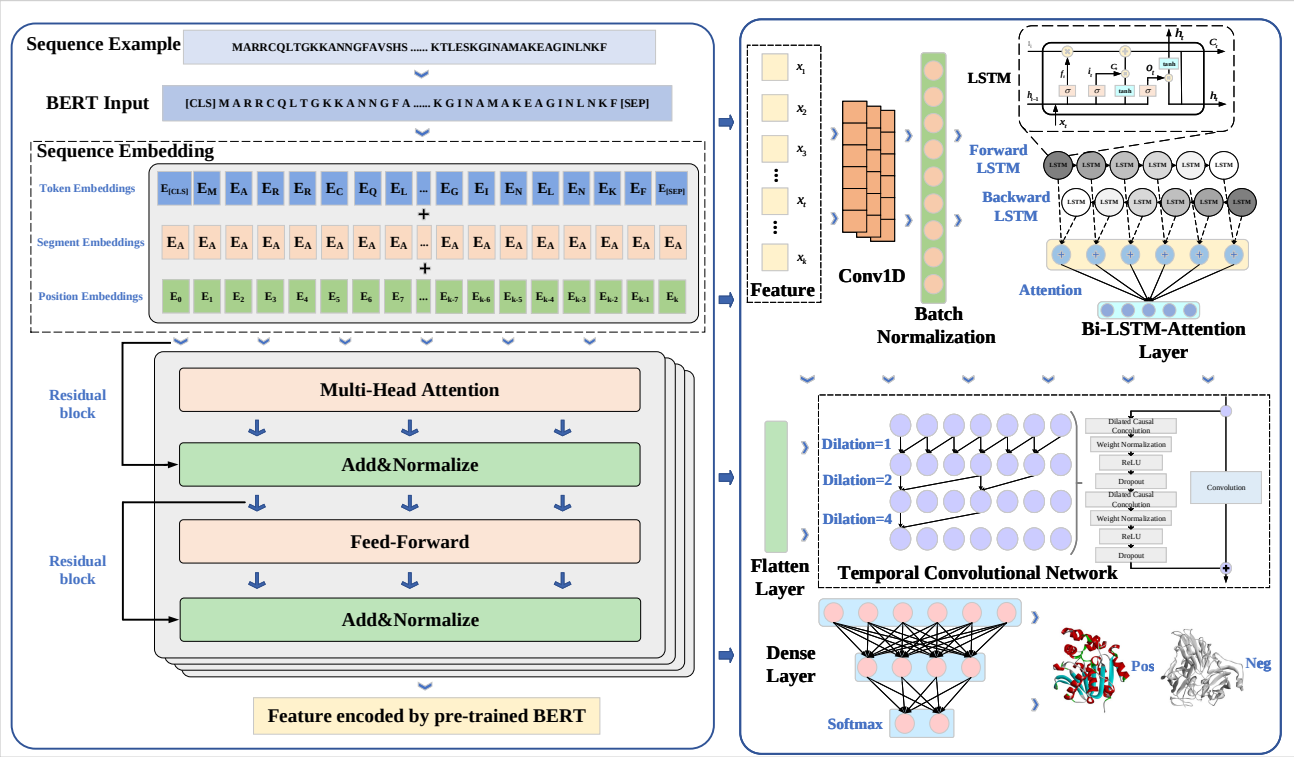

Fig. 1. Overall framework of BBATProt

processing tasks through a deep understanding of context and semantic relationships. In contrast to traditional recurrent neural networks and long short-term memory (LSTM) models, the architecture in this study mitigates performance degradation caused by long-term dependencies, enhancing both parallel computation and the capture of long-distance information. An innovation on this architecture, BERT, employs a bidirectional Transformer encoder that fully considers contextual information within the input.

In this study, BERT is leveraged to map each amino acid sequence into a feature vector, treating each molecular sequence in the reference dataset as a sentence and each amino acid as a word. The multi-head self-attention mechanism employed by BERT captures intricate relationships between every possible pair of amino acids, thereby enhancing feature representation. By utilizing transfer learning, BERT effectively applies insights from natural language processing to protein sequences, significantly reducing the reliance on extensive task-specific data while facilitating efficient feature extraction even in data-scarce scenarios. The selected pretraining model for this experiment is BERT-Small, consisting of 4 encoder layers, each with 8 attention heads and 512 hidden units. This process encompasses two primary steps: sequence embedding and feature extraction through the encoder layers. During the sequence embedding phase, [CLS] and [SEP] tokens are incorporated to ensure proper embedding within the BERT framework. As BERT processes each amino acid, it generates token, segment, and position embeddings, enabling the model to comprehend relationships among different amino acids in the sequence. Token embeddings represent words based on a specific tokenization methodology, while segment embeddings distinguish positions within the sequence. Position embeddings convey both relative and absolute positional information, as articulated in formulas (1) and (2), following the position vector framework proposed by Vaswani et al. (32).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{1}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{2}$$

At the encoder layer stage, multi-head self-attention and feedforward neural networks are employed to process embedding vectors and transform them into more intricate feature representations. The multi-head self-attention sub-layer performs attention calculations at each position of every input sequence, capturing the relationships between the current position and all other positions.

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

$$\text{head}_i = \text{Attention}(XW_i{}^Q, XW_i{}^K, XW_i{}^V) \tag{4}$$

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{5}$$

In the above formula, Q, K, and V, respectively, represent the matrices $Q(uery)$, $K(ey)$, and $V(alue)$ obtained through different linear transformations of the input sequence. $W_i{}^Q$,$W_i{}^K$,$W_i{}^V$ and $W_i{}^O$ represent weight matrices, where

d_k denotes the dimensionality of the key vectors in the Key matrix. The feedforward neural network (FFNN) sub-layer achieves information processing by applying nonlinear transformations to the contextual representations at each position. This sub-layer consists of an activation function, like $ReLU$, along with two linear transformations, facilitating the efficient extraction of features. The specific formula is as follows:

$$FFNN(X) = ReLu(XW_1 + b_1)W_2 + b_2 \tag{6}$$

where W_1 and W_2 stand for weight parameters, b_1 and b_2 signify the bias terms. Additionally, X indicates the representation obtained after passing through the multi-head self-attention sub-layer.

Network architecture post feature extraction

In this study, we constructed a function prediction network framework that integrates CNN, Bi-LSTM, attention mechanisms, and TCN. After protein sequences are encoded using BERT, the framework is employed to further enhance feature extraction. The CNN autonomously extracts feature information through one-dimensional convolutional layers, refining the data initially obtained from BERT. The outputs from the CNN are then directed to the Bi-LSTM layer for further processing.

The LSTM structure effectively captures long-range associations in protein sequences by incorporating memory cells and forget gate mechanisms. Memory cells retain essential information while the forget gates discard irrelevant features, thus improving the extraction of complex structures. Every Bi-LSTM unit accurately captures the temporal dependencies within the sequence by processing features both forward and backward. Given that protein sequences exert functional influences from both the N- and C-terminal ends, Bi-LSTM is particularly advantageous for tasks such as amino acid sequence prediction, as it can simultaneously extract critical information from both ends.

The attention mechanism enhances the model's flexibility, allowing it to balance and selectively focus on different segments of the input when analyzing protein sequences. Key components include queries, keys, values, attention scores, softmax functions, and final attention outputs. The comparison between the query and key is computed, the results are normalized using the softmax function, and the normalized attention weights are applied to the values to produce the outputs that are ultimately required. In this context, BBATProt leverages self-attention to dynamically adjust the significance of each region in the sequence, thereby enhancing the model's explainability. This approach clarifies how various parts of the protein sequence contribute to the final prediction, rendering the decision-making process more interpretable compared to conventional black-box models.

Following this, the flatten layer transforms the multidimensional input data generated by the attention mechanism into a one-dimensional array, which is then passed to the TCN layer, effectively connecting features across different layers and enhancing the model's ability to represent the input data.

The core components of the TCN comprise one-dimensional convolutional layers, residual connections, dilated convolutions, and causal convolutions. The incorporation of residual and causal connections in TCNs prevents information leakage from future positions, maintaining a clear cause-and-effect relationship in sequence processing. The temporal relationships and positional relation yeilded from procession, reflecting the protein folding and conformational information, are critical for function . The model's receptive field grows exponentially as the number of layers increases thanks to the dilated convolutions, facilitating efficient capture of global sequence dependencies. This design allows for a clearer trace of which segments of the sequence influence the output, thereby enhancing interpretability compared to other recurrent models.

The integration of Bi-LSTM and TCN in BBATProt is particularly advantageous because these architectures complement each other's strengths in local and global feature extraction. While Bi-LSTM excels in capturing sequential and bidirectional dependencies, it can face challenges with parallelization and efficient processing of lengthy sequences. Conversely, TCN is adept at capturing long-range dependencies and enabling parallel processing, though it lacks the bidirectional insight provided by Bi-LSTM. By combining these two architectures, BBATProt effectively captures both local sequential information and long-range dependencies. This framework offers a more transparent perspective on how features from various segments of the sequence are processed and synthesized to produce the final prediction. This transparency is vital in biological applications, where comprehending the underlying mechanics takes precedence over merely achieving high predicted accuracy.

Ultimately, in the Dense layer, feature information from each layer is integrated, with dimensions progressively reduced in a hierarchical manner. The final classification output is achieved using a sigmoid activation function, ensuring that the features learned by the network are distilled into a concise result.

Model Evaluation Parameters

Six metrics were taken into consideration in order to assess the performance of the classification models: F1 score (FSc), specificity (SPE), sensitivity (SEN), precision (PRE), accuracy (ACC), and Matthews correlation coefficient (MCC). These measures have the following definitions:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \tag{8}$$

$$SEN = \frac{TP}{TP + FN} \tag{9}$$

$$SPE = \frac{TN}{TN + FP} \tag{10}$$

$$PRE = \frac{TP}{TP + FP} \tag{11}$$

$$FSc = \frac{2 \times SEN \times PRE}{SEN + PRE} \tag{12}$$

In the above formula(7)-(12), TP and TN designate the counts of correctly forecasted positive and negative instances respectively, while FP and FN denote the quantities of erroneously predicted positive and negative instances.

Table 2. Comparative Representational Ability of BERT Encoding Features Across Different Sizes (%).

Pre-trained model	ACC	MCC	SEN	SPE	PRE	FSc	AUROC
BERT-Tiny	92.74	85.97	90.00	**95.49**	**95.44**	92.36	92.9
BERT-Mini	93.20	86.60	92.84	93.56	93.70	93.17	92.7
BERT-Small	**94.21**	**88.54**	**93.48**	94.94	94.96	**94.15**	**94.21**
BERT-Medium	93.07	86.31	92.24	93.91	93.95	93.01	93.10
BERT-Base	89.55	79.71	89.11	90.00	90.77	89.62	90.20
BERT-Protein	86.94	74.67	87.33	86.55	87.63	87.02	87.20

Note: The optimal performance for each metric is indicated in bold.

Experimental Analysis

To comprehensively validate the universal applicability of the BBATProt framework in peptide and protein function prediction tasks, this study focused on three key categories: enzyme function prediction, peptide function prediction, and PTM site prediction. Five diverse datasets were utilized to ensure a robust evaluation of the model's versatility, covering the prediction of various biological functions, including carboxyl ester hydrolytic activity, antimicrobial activity, dipeptidyl peptidase-lV inhibition, and potential sites for post-translational modifications such as lysine crotonylation (Kcr) and lysine glycation (Kgly).

Comparison of BERT Representational Capability Across Different Sizes

Google Research has recently released pre-trained BERT models in multiple sizes (14). Meanwhile, Zhang et al. selected 556,603 protein data entries from UniProt to pretrain their model (33). In this study, through detailed comparative experiments, the goal is to determine the most suitable feature representation. Six representative models were selected for fine-tuning out of all the pre-trained BERT models: BERT-Tiny, BERT-Mini, BERT-Small, BERT-Base, and BERT-Protein. Specific parameters can be found in the supplementary data, and the comparative results are presented in the charts 2. Relatively, the Small-sized models show better results during training. Consequently, the pre-trained BERT-Small model was chosen to extract characteristics from protein sequence data.

t-SNE Visualization of Features Extracted Layer by Layer by BBATProt on Protein Dataset

The core task of deep learning in sample classification is the continuous extraction of features with unique representations, a process that evolves throughout the entire network structure. To thoroughly depict the advancement of the BBATProt model in the domain of feature learning, the concept of t-distributed stochastic neighbor embedding (t-SNE) has been brought into light for the reduction of dimensionality and visualization of information prevalent in the samples (34). This particular pathway paves the way to a heightened understanding of the abstraction and learning of sample features across distinct network strata. To elucidate the hierarchical feature extraction within the network, the analysis progresses from the raw samples through the convolutional layer, Bi-LSTM layer, attention layer, and TCN layer, culminating in the final Dense layer.

In the visualization graph shown in Fig. 2, positive samples are represented by blue dots and negative samples by green dots. A significant observation from this experiment is that with each additional layer in the neural network structure, the deeper layers exhibit enhanced capabilities for feature extraction.

Comparison of BBATProt with existing predictors on hydrolytic enzyme datasets

This segment starts with the formulation of a benchmark dataset containing carboxylesterases identified by EC number 3.1.1.1. Known for their ability to act as catalysts, these enzymes facilitate the hydrolysis of ester. Beyond their functions in biological processes like drug metabolism, food digestion, and xenobiotic detoxification, these enzymes play a crucial role in various industries, including renewable energy, food processing, and bioremediation in environmental engineering. Given the extensive time and effort required by traditional experimental methods, computational approaches offer valuable and efficient alternatives.

The positive dataset was obtained from the UniProt. In order to improve feature extraction, sequences exceeding a length of 512 were excluded, focusing on those within the range of 0-512. Furthermore, to reduce the influence of high sequence similarity on the experimental results, CD-HIT was utilized to filter out redundant sequences with a similarity threshold above 80%. Negative samples were randomly generated from Swiss-Prot, maintaining the same length range as positive samples. These negative samples also underwent CD-HIT processing with a threshold set at 30%. Ultimately, a training dataset comprising 6960 hydrolytic enzyme samples was obtained, with 3480 each for positive and negative samples.

To demonstrate the effectiveness of the proposed network model, a comparison was drawn against various conventional machine learning classification strategies, including KNN, RF, SVM, and XGBoost. Hydrolytic enzymes were selected specifically for a 10-fold cross-validation, the comprehensive results of which are presented in Table 3. As these data reveal, BBATProt excels in the intricate domain of enzyme protein function prediction, exhibiting peak performance across a range of evaluation standards.

Additionally, a comparative analysis was conducted to evaluate the encoding effectiveness of two NLP models, BERT and Word2Vec (35), on the hydrolase dataset. Both models were applied to encode the data, and their performance was assessed using an identical neural network architecture for classification. Word2Vec uses a dual-modality that encompasses both skip-gram and CBOW strategies. Continuous Bag of Words (CBOW) mainly concentrates on forecasting the desired term when provided with contextual clues, whereas the skip-gram tactic emphasizes the estimation of the surrounding wording based on the specific target term. Table 4 shows the efficacy measures of each system via a 10-fold cross-validation.

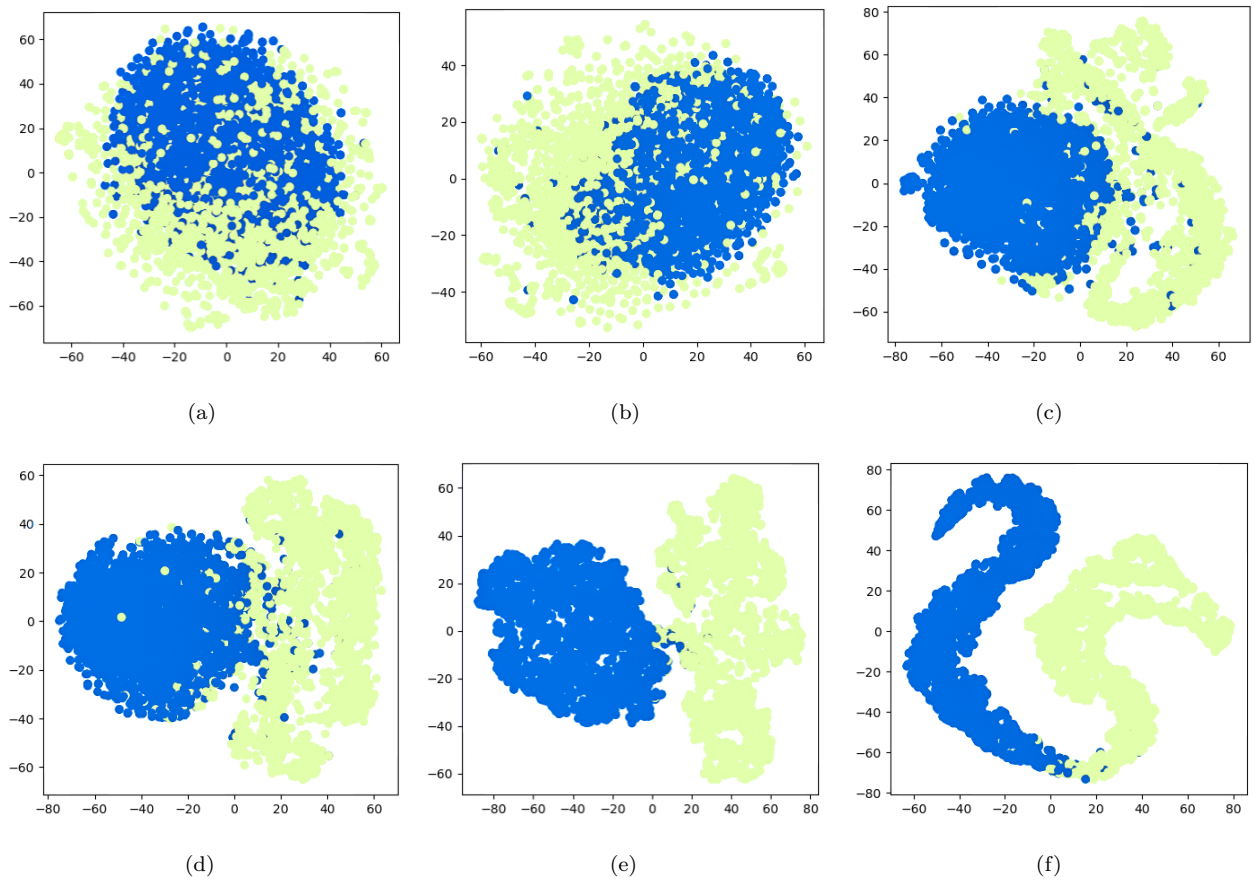

Fig. 2. The t-SNE visualization of features extracted layer by layer through BBATProt on the protein dataset. (a) t-SNE visualization of encoded features after BERT encoding. (b) t-SNE visualization of features after the convolutional layer. (c) t-SNE visualization of features after the Bi-LSTM layer. (d) t-SNE visualization of features after the Attention layer. (e) t-SNE visualization of features after the TCN layer. (f) t-SNE visualization of features after the Dense layer.

Table 3. 10-Fold Cross-Validation Comparison Between BBATProt and Other Machine Learning Methods on the Hydrolysis Enzyme Dataset with the Same BERT Encoding Features(%).

Predictor	ACC	MCC	SEN	SPE	PRE	FSc	AUROC
BBATProt	**96.60**	**93.33**	**95.59**	**97.79**	**97.97**	**96.71**	96.63
BERT+RF	75.57	51.30	71.50	79.65	77.87	74.53	97.97
BERT+XGBOOST	80.63	61.28	79.25	82.02	81.51	80.34	**99.58**
BERT+KNN	79.35	59.87	88.92	69.88	74.66	81.13	86.35
BERT+SVM	83.79	67.61	82.38	85.23	84.78	83.54	84.60

Note: The optimal performance for each metric is indicated in bold.

The outcome decisively attests that the pre-trained BERT model is superior to the other two NLP models in every performance measurement, strongly underscoring the notable effectiveness of BERT encoding compared with Word2Vec.

Comparison of BBATProt with existing predictors on peptide datasets

The assessment of BBATProt's predictive capability on peptide datasets necessitates a comprehensive examination of both the AMP and DPP-IV inhibitory peptide datasets.

Comparison of BBATProt with existing predictors on antimicrobial peptide datasets

In this section, we provide a detailed comparison of BBATProt with 13 SOTA methods (36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 26) using the XUAMP dataset. This comparison was conducted through a comprehensive evaluation in an independent testing environment. The results demonstrate that BBATProt significantly outperforms the other methods across key metrics, including ACC, MCC, SEN, Fsc, and the AUROC. As shown in Table 5, BBATProt shows an improvement ranging from 2.81% to 31.96% in ACC, 5.42% to 63.68% in MCC, 0.56% to 64.4% in SEN, 2.39% to 49.46% in FSc, and 0.71% to 40.51% in AUROC when compared to its counterparts.

Table 4. Comparative Results of 10-Fold Cross-Validation on the Hydrolysis Enzyme Dataset Using Different NLP-based Feature Representation Models(%).

Predictor	ACC	MCC	SEN	SPE	PRE	FSc	AUROC
BBATProt	**96.60**	**93.33**	**95.59**	**97.79**	**97.97**	**96.71**	**96.63**
Word2Vec-CBOW+Conv+BiLSTM+Att+TCN+Dense	48.36	NA	50.00	50.00	24.18	32.59	48.36
Word2Vec-Skip-gram+Conv+BiLSTM+Att+TCN+Dense	48.22	NA	60.00	40.00	29.11	39.19	48.22

Note: The optimal performance for each metric is indicated in bold. NA is used as a representation when the result is not available.

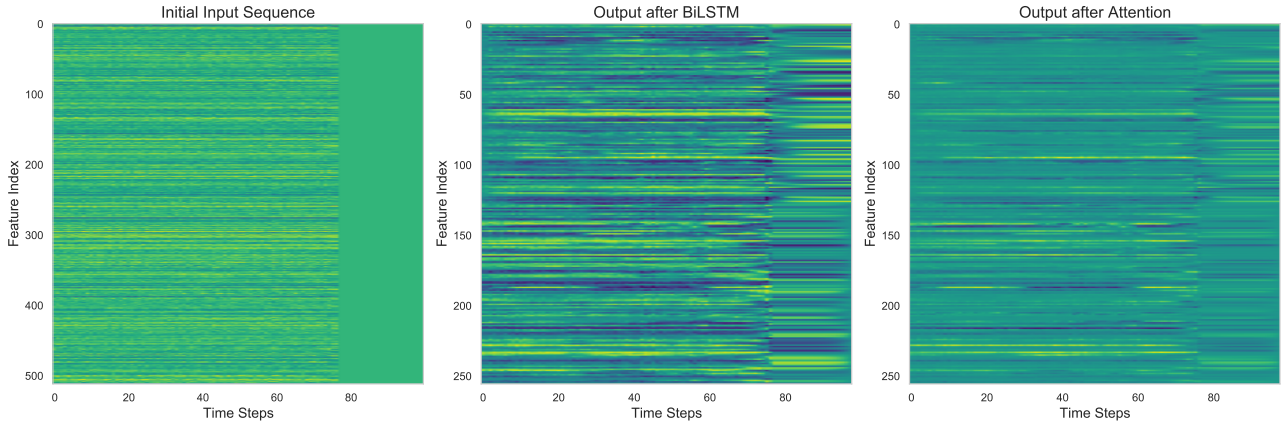

Fig. 3. Visualization of the attention mechanism.

Additionally, BBATProt ranked 6th and 2nd in the SPE and PRE evaluations, respectively. While AMPEP achieved the highest scores in these two metrics, its overall average performance across the other five evaluation metrics was 32.80% lower than that of BBATProt. This indicates that BBATProt provides a more comprehensive extraction of feature information, emphasizing its strengths across various predictive tasks and suggesting potential avenues for further improvement in protein analysis. By examining the model characteristics of AMPEP in more detail, valuable insights can be gained for optimizing BBATProt in future studies. Moreover, the visualization of the attention mechanism, as shown in Figure 4, reveals that BBATProt effectively filters out irrelevant information during the computation process, demonstrating a focused and refined approach to data analysis. This visual evidence supports the conclusion that the attention mechanism is crucial for enhancing BBATProt's ability to accurately predict protein functions by concentrating on the most relevant features.

To comprehensively assess the effectiveness of BBATProt in AMP prediction, a benchmark dataset was constructed. During dataset construction, sequences with more than 20 natural standard amino acids were excluded to ensure data consistency. The AMP dataset was compiled from various sources, including the AMPer, APD3, and ADAM databases (35; 46; 47). Data cleaning involved setting a CD-Hit threshold of 90% to remove redundant sequence information. Correspondingly, non-AMP data were sourced from UniProt, with a constraint on the residue length of protein fragments between 5 and 100 to maintain sequence lengths similar to those of the AMP dataset for experimental validity. The CD-Hit redundancy removal threshold for the non-AMP dataset was set to 40%. In addition, sequences with annotations containing terms such as 'Defensin', 'Antimicrobial', 'Antibiotic' or 'Antifungal' were excluded to ensure data purity.

Using the benchmark dataset, multiple repetitions of 10-fold cross-validation were performed to accurately evaluate the performance and generalization capability of the predictor. The resulting averages for each validation and across five repetitions are detailed in Figure . This method ensures a robust assessment of the performance of the model on various dataset subsets, emphasizing the reliability and repeatability of the results.

Comparison of BBATProt with existing predictors on inhibitory peptide datasets

Due to their crucial role in pharmaceutical development and diabetes treatment, distinguishing between inhibitory and non-inhibitory peptides of dipeptidyl peptidase-IV (DPP-IV) has become a key focus. The BBATProt model was trained and evaluated on a dataset designed to differentiate these two classes, including both inhibitory DPP-IV peptides and non-inhibitory counterparts (27). The dataset was split into a benchmark set and an independent test set, each containing an equal distribution of both peptide types. Specifically, the benchmark set comprised 532 samples, while the independent set contained 133 samples. As presented in Table 6, BBATProt outperforms other models in terms of ACC, MCC, SEN, and SPE, and shows comparable AUROC performance to the current SOTA models.

Comparison of BBATProt with existing predictors on PTM site prediction datasets

Concerning PTM prediction challenges such as protein lysine crotonylation and protein lysine glycation, these alterations can affect the functionality, stability, and affinity of proteins,

Table 5. Performance Comparison Results of BBATProt and Existing Methods on Independent AMP Dataset(%).

Predictor	ACC	MCC	SEN	SPE	PRE	FSc	AUROC
BBATProt	**85.96**	**71.98**	**88.20**	83.71	84.41	**86.26**	**93.81**
ADAM-HMM	68.40	39.00	52.10	84.70	77.30	62.30	68.40
ADAM-SVM	61.20	26.40	34.60	87.80	73.90	47.10	61.20
AmpGram	56.40	13.10	44.50	68.20	58.40	50.50	54.70
AMAP	60.20	25.00	31.40	89.10	74.20	44.10	60.20
AMPEP	65.80	42.50	32.50	**99.20**	**97.50**	48.70	72.70
AMPfun	67.40	41.40	40.60	94.30	87.70	55.50	73.50
AMPScannerV2	56.80	13.70	52.30	61.30	57.50	54.80	58.50
APIN	57.90	16.30	44.60	71.20	60.70	51.40	57.50
APIN-fusion	56.00	12.30	45.70	66.30	57.60	51.00	55.40
iAMP-Attenpred	83.15	66.56	87.64	78.65	80.41	83.87	93.10
iAMPpred	54.00	8.30	65.20	42.80	53.30	58.70	57.50
Deep-AMPEP30	55.80	11.80	45.70	65.80	57.20	50.80	53.30
sAMPpred-GAT	71.50	46.40	53.00	90.00	NA	NA	77.70

Note: The optimal performance for each metric is indicated in bold. NA is used as a representation when the result is not available.

Table 6. Performance Comparison Results of BBATProt and Existing Methods on the DPP-IV Inhibitory Peptide Dataset(%).

Dataset	Predictor	ACC	MCC	SEN	SPE	PRE	FSc	AUROC
Benchmark	BBATProt	**95.6**	**91.64**	**94.23**	**97.69**	**93.68**	**95.75**	**95.58**
	iDPPIV	81.9	64.3	NA	81.9	81.8	NA	87
	iDPPIV-SCM	85.8	71.7	NA	87.7	83.9	NA	94
	IPPF-FE	NA	NA	NA	NA	NA	NA	NA
independent-data	BBATProt	**89.1**	**78.25**	**90.62**	**87.22**	**90.98**	**88.89**	93.04
	iDPPIV	79.7	59.4	NA	78.9	80.5	NA	84.7
	iDPPIV-SCM	86.5	73.1	NA	87.4	85.60	NA	93.90
	IPPF-FE	86.64	73.7	NA	85.71	87.97	NA	**94.25**

Note: The optimal performance for each metric is indicated in bold. NA is used as a representation when the result is not available.

subsequently steering their biological operations. However, experimental methods for site prediction are both expensive and time-consuming. In contrast, computational methods can provide reasonable predictions in a highly efficient and cost-effective manner. Using bioinformatics tools and algorithms, protein sequences and structures can be analyzed to predict potential modification sites.

To validate the advanced nature of the designed network structure, BBATProt was compared with some classical machine learning classification methods, including KNN, RF, SVM, and XGBoost. A 10-fold cross-validation was implemented for Kcr site prediction, with the outcomes detailed in Table 7. The evidence indicates the superior effectiveness of BBATProt, most notably in brief site prediction duties. This can be attributed to the fact that BBATProt is based on hierarchical learning, which allows it to adapt to protein data of different lengths and extract high-level features, making the model more effective in handling complex classification tasks.

To avoid overestimating the performance of BBATProt, a comparison was made between its training results and those of other existing SOTA models on independent test sets for both Kcr and Kgly site prediction problems. In Table 5, BBATProt demonstrates superior performance across almost all metrics compared with the other models specifically designed for Kcr site prediction (28; 50; 51). Although it lags slightly behind the CKSAPP_CrotSite algorithm in terms of SEN, this may be attributed to the lack of specific parameter tuning in BBATProt for addressing Kcr site prediction issues. This conclusion provides a potential avenue for future improvement research. Notably, in terms of ACC, MCC, SPE, PRE, FSc, and AUROC, BBATProt exhibits improvements of 8.94%, 16.91%, 25.56%, 5.3%, and 9.82%, respectively, compared with the CKSAPP_CrotSite algorithm model. This underscores the significant overall performance advantage of BBATProt.

In the Kgly site prediction task presented in Figure 6, BBATProt demonstrates a clear superiority over other advanced SOTA predictors across all evaluated metrics (29; 52; 20; 53). Notably, for the Sensitivity (SEN) metric, BBATProt shows an improvement range of 18.9% to 56.9% compared to other predictors, underscoring its robust capability to accurately capture positive samples. This performance highlights the significant feasibility of BBATProt in post-translational modification (PTM) site prediction, emphasizing its effectiveness in identifying relevant sites with high accuracy.

Conclusion

The function prediction of protein and peptide based on amino acid sequences is crucial in both academic research and industrial applications. This study introduces BBATProt, an innovative framework designed to tackle this challenge. Diverse biological datasets were transformed into peptide and protein datasets with varying amino acid sequence lengths, targeting prediction tasks such as hydrolases, AMP, DPP-IV inhibitory peptides, and PTM at Kcr and Kgly sites.

Table 7. 10-Fold Cross-Validation Comparison Between BBATProt and Other Machine Learning Methods on the Kcr Site Prediction Dataset with Consistent BERT Encoding Features(%).

Predictor	ACC	MCC	SEN	SPE	PRE	FSc	AUROC
BBATProt	**94.95**	**89.93**	**94.05**	**95.86**	**95.80**	**94.91**	94.95
BERT+RF	74.34	49.49	83.36	65.34	70.63	76.46	96 01
BERT+XGBOOST	76.86	54.10	82.80	70.91	74.00	78.15	**97.38**
BERT+KNN	70.12	40.23	69.48	70.74	70.36	69.91	78.70
BERT+SVM	79.47	59.30	84.97	73.97	76.56	80.53	92.40

Note: The optimal performance for each metric is indicated in bold.

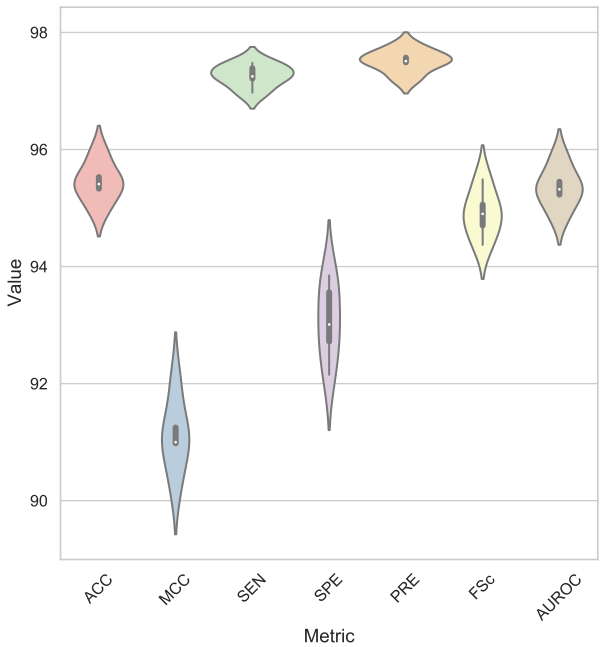

Fig. 4. Five times and average performance results of 10-fold cross validation method based on the AMP benchmark dataset(%).

Comprehensive evaluations across multiple datasets confirm its superior accuracy, robustness, and generalization capabilities compared to SOTA models. Furthermore, this study enhances the interpretability of what is typically a black-box model by leveraging visualization techniques. The sequential network architecture, built around the spatial conformations of proteins, further validates the model's effectiveness. Additionally, enhanced feature extraction through transfer learning has proven instrumental in improving prediction performance. Together, these approaches ensure that BBATProt not only excels in predictive power but also provides a more explainable and transparent model, offering deeper insights into how specific features of amino acid sequences contribute to functional predictions.

Although BBATProt performs well in the function prediction of protein and peptide, there is still room for improvement. Future research will focus on feature integration, further optimizing prediction performance by incorporating structural information of peptides or proteins into the integrated model's input. Additionally, plans include the development of an intuitive and user-friendly web server to offer public prediction services, enhancing accessibility to a wider researcher base and expanding practical applications.

Acknowledgments

This work was supported in part by the National Science Fund for Distinguished Young Scholars of China under Grant 62225303; in part by the National Natural Science Funds of China under Grant 62303039, 32371325 and 62433004; in part by the China postdoctoral Science Foundation BX20230034 and 2023M730190; in part by the Fundamental Research Funds for the Central Universities buctrc202201, QNTD2023-01; in part by Beijing Natural Science Foundation (L241014) and in part by the High Performance Computing Platform, College of Information Science and Technology, Beijing University of Chemical Technology.

Data Availability

Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/Xukai-YE/BBATProt.

Key Ponits

- BBATProt proposes an explainable neural network framework for protein function prediction.
- The model outperforms state-of-the-art methods in various protein and peptide prediction tasks.
- Visualization confirms model interpretability, highlighting attentions focus on relevant data.

References

1. Muhammad Attique, Muhammad Salman Farooq, Anouar Khelifi, et al. Prediction of therapeutic peptides using machine learning: computational models, datasets, and feature encodings. *IEEE Access*, 11(8):148570–148594, 2020.
2. Vladimir Gligorijević, Patrick D Renfrew, Tomasz Kosciolek, et al. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12:3168, 2021.
3. John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.
4. Alex L Mitchell, Teresa K Attwood, Patricia C Babbitt, et al. InterPro in 2019: improving coverage, classification

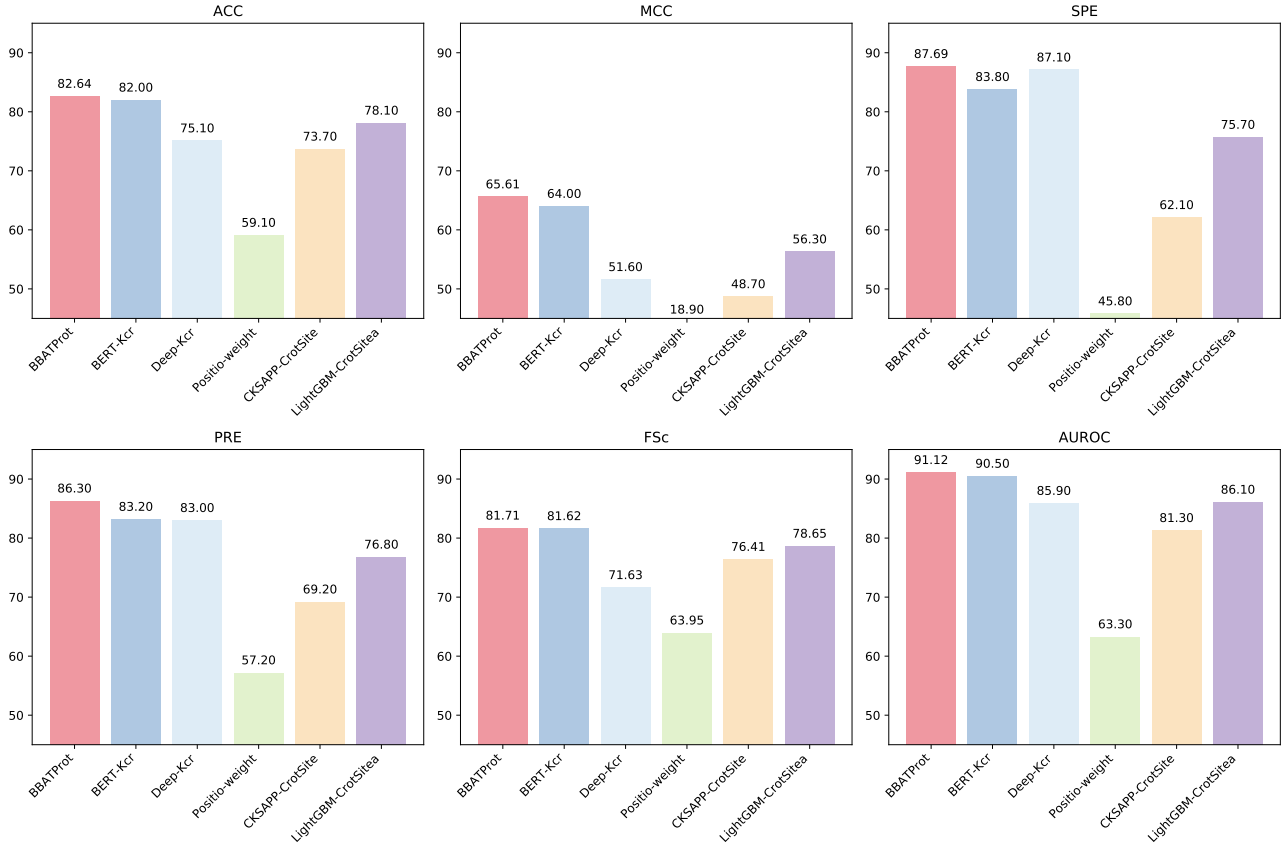

Fig. 5. Performance Comparison Results of BBATProt and Existing Methods on Independent Kcr Site Prediction Dataset(%).

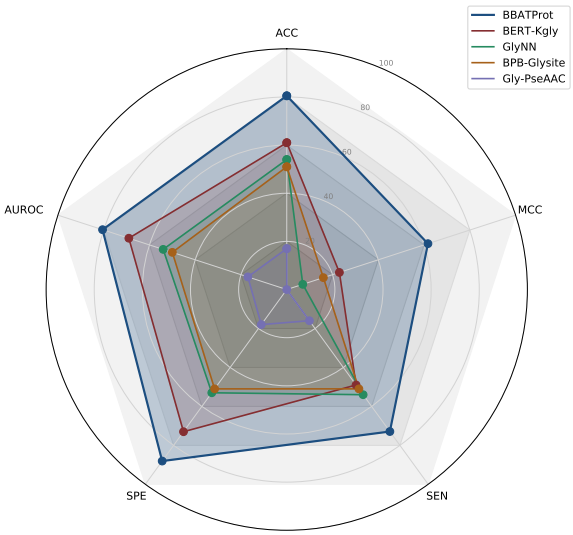

Fig. 6. Performance Comparison Results of BBATProt and Existing Methods on the Kgly Site Prediction Dataset.(%).

and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360, 2019.

5. Boqiao Lai, Jinbo Xu. Accurate protein function prediction via graph attention networks with predicted structure information. *Briefings in Bioinformatics*, 23(1):bbab502, 2022.
Briefings in Bioinformatics, 23(1):bbab502, 2022.
6. Han Yu, Xiaozhou Luo. IPPF-FE: An integrated peptide and protein function prediction framework based on fused features and ensemble models. *Briefings in Bioinformatics*, 24(1):bbac476, 2023.
7. Thomas M. Cover, Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.
8. Leo Breiman. Random forests. *Machine Learning*, 36:105–139, 2001.
9. Vladimir N Vapnik. The nature of statistical learning theory. *IEEE Transactions on Neural Networks*, 8(6):1564, 1997.
10. Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.
11. Nicolas Sapoval, Ali Aghazadeh, Michael G Nute, et al. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13:1728, 2022.

12. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, June 2017.
13. Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 2015)*, 2015.
14. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.
15. Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *Learning*, 2018.
16. Alexander Rives, Joshua Meier, Tom Sercu, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15):1–12, 2021.
17. Jaspreet Singh, Kuldip Paliwal, Thomas Litfin, et al. Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. *Scientific Reports*, 12(1):7607, 2022.
18. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112-7127, 2022.
19. Zhang Yue, Lin Jianyuan, Zhao Lianmin, et al. A novel antibacterial peptide recognition algorithm based on BERT. *Briefings in Bioinformatics*, 22(6):bbab200, 2021.
20. Jing Xu, Fuyi Li, André Leier, et al. Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Briefings in Bioinformatics*, 22(5):bbab083, 2021.
21. Hongliang Zou, Zhijian Yin. Identifying Dipeptidyl Peptidase-IV Inhibitory Peptides Based on Correlation Information of Physicochemical Properties. *International Journal of Peptide Research and Therapeutics*, 27(4):2651–2659, 2021.
22. Dandan Wang, Liwei Zou, Qiang Jin, et al. Human carboxylesterases: a comprehensive review. *Acta Pharmaceutica Sinica B*, 8(5):699–712, 2018.
23. Hao Lv, Fu-Ying Dao, Zheng-Xing Guan, et al. Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. *Briefings in Bioinformatics*, 22(4):1–10, 2021.
24. Ying Huang, Beifang Niu, Ying Gao, et al. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.
25. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43, D204–D212, 2015.
26. Yan, Ke, Hongwu Lv, Yichen Guo, et al. sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. *Bioinformatics*, 39(1):1–8, 2023.
27. Phasit Charoenkwan, Chanin Nantasenamat, Md Mehedi Hasan, et al. iBitter-Fuse: a novel sequence-based bitter peptide predictor by fusing multi-view features. *International Journal of Molecular Sciences*, 22(16):8958, 2021.
28. Yun Qiao, Xue Zhu, Hui Gong. BERT-Kcr: Prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics* , 38(3):648–654, 2022.
29. Yinbo Liu, Yufeng Liu, Gang-Ao Wang, et al. BERT-Kgly: A Bidirectional Encoder Representations From Transformers (BERT)-Based Model for Predicting Lysine Glycation Site for Homo sapiens. *Frontiers in Bioinformatics*, 2:834153, 2022.
30. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
31. Jialin Yu, Shaoping Shi, Fang Zhang, et al. PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization. *Bioinformatics*, 35(16):2749–2756, 2019.
32. Ashish Vaswani, Noam Shazeer, and Niki Parmar. Attention is all you need. In *Proceedings of 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.
33. Yue Zhang, Jianyuan Lin, Lianmin Zhao, et al. A novel antibacterial peptide recognition algorithm based on BERT. *Briefings in Bioinformatics*, 22(6):bbab200, 2021.
34. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.
35. Tomas Mikolov, Kai Chen, Greg Corrado, et al. Efficient estimation of word representations in vector space. *Computer Science*, 2013.
36. Hao-Ting Lee, Chen-Che Lee, Je-Ruei Yang, et al. A large-scale structural classification of antimicrobial peptides. *BioMed Research International*, 2015:475062, 2015.
37. Michał Burdukiewicz, Katarzyna Sidorczuk, Dominik Rafacz, et al. Proteomic screening for prediction and design of antimicrobial peptides with AmpGram. *International Journal of Molecular Sciences*, 21(12):E4310, 2020.
38. Sadaf Gull, Nauman Shamim, Fayyaz Minhas. AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides. *Computers in Biology and Medicine*, 107:172–181, 2019.
39. Pratiti Bhadra, Jielu Yan, Jinyan Li, et al. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Scientific Reports*, 8(1):1697, 2018.
40. Chia-Ru Chung, Ting-Rung Kuo, Li-Ching Wu, et al. Characterization and identification of antimicrobial peptides with different functional activities. *Briefings in Bioinformatics*, 21(3):1098, 2020.
41. Daniel Veltri, Uday Kamath, Amarda Shehu. Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 34(16):2740–2747, 2018.
42. Xin Su, Jing Xu, Yanbin Yin, Xiongwen Quan, Han Zhang. Antimicrobial peptide identification using multi-scale convolutional network. *BMC Bioinformatics*, 20(1):730, 2019.
43. Wenxuan Xing, Jie Zhang, Chen Li, et al. iAMP-Attenpred: a novel antimicrobial peptide predictor based on BERT feature extraction method and CNN-BiLSTM-Attention combination model. *Briefings in Bioinformatics*, 25(1):bbad443, 2023.

44. Prabina Kumar Meher, Tanmaya Kumar Sahu, Varsha Saini, et al. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific Reports*, 7:42362, 2017.
45. Jielu Yan, Pratiti Bhadra, Ang Li, et al. Deep-AmPEP30: Improve short antimicrobial peptides prediction with deep learning. *Molecular Therapy - Nucleic Acids*, 20:882–894, 2020.
46. Christopher D. Fjell, Robert E. W. Hancock, Artem Cherkasov. AMPer: a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics*, 23(9):1148–1155, 2007.
47. Guangshun Wang, Xia Li, Zhe Wang. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, 44(1):D1087–D1093, 2016.
48. Hongliang Zou, Zhijian Yin. Identifying Dipeptidyl Peptidase-IV inhibitory peptides based on correlation information of physicochemical properties. *International Journal of Peptide Research and Therapeutics*, 27(4):2651–2659, 2021.
49. Phasit Charoenkwan, Sakawrat Kanthawong, Chanin Nantasenamat, et al. iDPPIV-SCM: a sequence-based predictor for identifying and analyzing dipeptidyl peptidase IV (DPP-IV) inhibitory peptides using a scoring card method. *Journal of Proteome Research*, 19(10):4125–4136, 2020.
50. Zhe Ju, Jian-Jun He. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC. *Journal of Molecular Graphics Modelling*, 77:200–204, 2017.
51. Yaning Liu, Zhaomin Yu, Cheng Chen, Yu Han, Bin Yu. Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net. *Analytical Biochemistry*, 609:113903, 2020.
52. Morten Bo Johansen, Lars Kiemer, Søren Brunak. Analysis and prediction of mammalian protein glycation. *Glycobiology*, 16(9):844–853, 2006.
53. Zhe Ju, Juhe Sun, Yanjie Li, Li Wang. Predicting lysine glycation sites using bi-profile bayes feature extraction. *Computational Biology and Chemistry*, 71:98–103, 2017.
54. Yan Xu, Li Li, Jun Ding, et al. Gly-PseAAC: Identifying protein lysine glycation through sequences. *Gene*, 602:1–7, 2017.