AGRI MAGAZINE
(International E-Magazine for Agricultural Articles)
Volume: 03, Issue: 08 (August, 2026)
Available online at http://www.agrimagazine.in

Expanding Peptide Therapeutics Through Computational Interaction Prediction

Parmila Bhukal, Anupama Roy, Sarika Jaiswal and *Mir Asif Iquebal
Division of Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012, India
***Corresponding Author's email: jiqubal@gmail.com**

Peptides are the small sequence of amino acids, usually contain fewer than 50 residues which play crucial role in various biological pathways including cellular signalling, programmed cell apoptosis, gene transcription, etc. Peptides perform their functions by interacting with different proteins. Inside the cellular environment, these interactions are approximately account for nearly ~15-40% of all protein-protein interactions (PPI). On the basis of molecular level, precise forecasting of these protein-peptide interactions plays crucial role in therapeutic peptide development and drug design. However, traditional lab methods were slow and expensive and earlier computational approaches are derived from sequences and structural features enhance the throughput but it doesn't have the ability to generalize among diverse proteins-peptide conformations. The Deep-ProBind model uses deep learning, it combines the ProtBERT model with evolutionary encodings (PsePSSM and DWT) to extract both sequence-based context and evolutionary based features. It also uses SHAP (SHapley Additive exPlanations) algorithm to improve interpretability, outperforms among multiple benchmark datasets, offering a scalable and generalizable solution for precise peptide-protein interactions prediction.

Keywords: Peptides, Deep learning, Transformer, SHAP, ProtBert, PsePSSM

Introduction

Peptides are the small sequence of amino acids, usually contain fewer than 50 residues which play important role in various biological processes such as cellular signalling, programmed cell apoptosis, gene transcription, etc. Peptides perform their functions by interacting with different proteins (Huang *et al.*, 2024). Unlike typical protein-protein interactions that involve large, structured interfaces, peptide-protein interactions frequently depend on short, linear motifs within the peptide sequence that bind transiently and specifically to surface pockets or domains of the protein receptor. It is estimated that such interactions make up ~15-40% of all protein-protein interactions in the cellular environment (Yin *et al.*, 2024). Inside the cell, enzymes work as biological catalysts which regulate many biochemical reactions, and finding ligands that can modulate their activity is important for improving disease diagnosis and therapy. Peptides can influence enzyme function, modify receptor responses and assist in transmembrane transport, providing opportunities for therapeutic interventions. These are structurally diverse, highly selective, and less toxic so there are fewer off-target effects compared to small molecule pharmaceuticals (Marqus *et al.*, 2017). Moreover, their capacity to mimic natural regulatory or signalling motifs allows them to bind to protein surfaces often considered 'undruggable' by conventional approaches (Fosgerau and Hoffmann, 2015). Despite these benefits, the development of peptide-based drugs has been mostly constrained by rapid degradation by proteases, short systemic half-life and poor oral bioavailability (Lau and Dunn, 2018). Recently, advancement in peptide synthesis, cyclization, backbone modification and formulation techniques has substantially enhance their pharmacokinetic

characteristics (Uhlig *et al.*, 2014). Study of molecular level of peptide-protein recognition play important role in rational peptide design. Publicly available structural databases including PeptiDB, PepBind, and PepX provide curated high-resolution protein-peptide complex data for studying interaction mechanisms and their associated relationship.

Various experimental methods including X-ray crystallography, nuclear magnetic resonance (NMR), yeast two-hybrid assays, and phage display utilized for identification of peptide-protein interactions. However, these experimental methods are resource-intensive, low-throughput and not feasible for screening large peptide libraries. The advent of deep learning, particularly natural language processing (NLP)-based models, has significantly advanced the field of peptide-protein interaction prediction. The Bi-directional Encoder algorithm, particularly efficient in this field based on Transformers (BERT) model, which is initially created for NLP purposes. It is able to capture contextual dependencies and model long-range interactions, making it suitable for analysis of peptide sequences where the functional behavior of a residue often depends on its surrounding sequence context. It can also be applied to the protein language models (PLM)like ProtBERT, which produce highly informative, high dimensional embeddings of peptides. These embeddings encode for structural, evolutionary, and biochemical properties allow for more precise and scalable predictions of peptide-binding behaviour. This data-driven, scalable approach offers an alternative to traditional laboratory-based approaches and greatly enhance the discovery of peptide-based therapeutics.

Computational algorithm-based prediction techniques increasingly fill this gap and offer scalable and cost-effective resolutions for *in silico* peptide screening. Artificial intelligence-based (AI)approaches including machine learning (ML) and deep learning (DL) methods can learn patterns from known instances to identify the sequence-level patterns which are associated with binding potential, thereby enables rapid identification of novel bioactive peptides. Transformers (BERT) model representations, which are initially created for NLP applications, have proven very effective in this domain. Among these approaches, Deep-ProBind combines transformer-based models of protein language with evolutionary features for the advancement of protein–peptide interactions (Khan *et al.*, 2025).

Model Architecture

Feature Encoding Schemes: Peptide sequences, short sequences of amino acids are represented by standard one-letter codes, act as model's raw input. These representative sequences must first be encoded in the numerical forms as machine learning based approaches work well with numerical data. Teh architecture of Deep-ProBind, proposed by (Khan *et al.*, 2025), which operates along two separate representation routes:

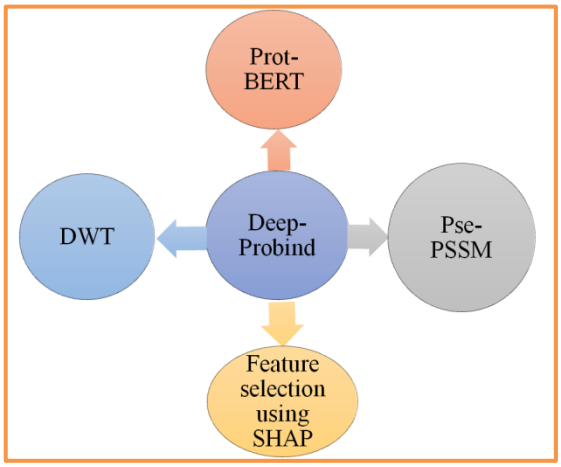

Figure.1: Overview of components of Deep-ProBind.

1. **Position-Specific Scoring Matrix (PSSM):** Position-Specific Iterated BLAST (PSI-BLAST), utilized each peptide as a query sequence aligns it against a curated protein sequence database. It creates a position-specific scoring matrix, an L×20 matrix (where L is the length of the peptide and 20 corresponds to the standard amino acids). For example, a peptide "ACDBG", resulting PSSM matrix will contain 5 rows (one for each residue) and 20 columns (amino acid substitution scores), encoding evolutionary substitution patterns at each position. In PSSM matrix, each element indicates the log-odds score of a particular amino acid at a specific sequence position based on homologous alignments.
2. **PsePSSM (Pseudo-PSSM):** PsePSSM utilized to convert the variable-length PSSM into a fixed-length vector, which includes global conservation patterns (mean values) and

sequence order information (correlations between positions). The average score per amino acid calculated which captures how each of the amino acid is retained throughout the peptide.

DWT-Discrete Wavelet Transform: DWT analyses the signals in the PSSM matrix and measures the conservation fluctuation across sequence positions. It breaks the signal into frequencies. It is a signal decomposition-based method utilized to break a 1-Dimesional signal (like a feature vector) into various frequency components such as low-frequency approximation and high-frequency approximation which captures the overall trend and local motifs, using low-pass and high pass filter, respectively.

ProtBERT: ProtBERT is a state-of-the-art protein language model design to accelerate the representations and biological insights of protein sequences (Elnaggar *et al.*, 2021). It employs Byte Pair Encoding (BPE), a subword tokenization algorithm extensively used in natural language processing. It overcomes the problems of fixed-length tokenization by identifying the biologically meaningful motifs and substructures in large protein databases such as UniRef50.

In this, positional embeddings which tells that ProtBERT's vocabulary contains pre-trained embedding vectors that are mapped to each token. These vectors are of size 768 or 1024. The token embeddings are supplemented with positional encodings. These encodings provide the details about each amino acid's location within the sequence. The complete input peptide turns into a form matrix,

$$\text{Input matrix } \epsilon R^{L \times D}$$

L = number of tokens

d = embedding dimension

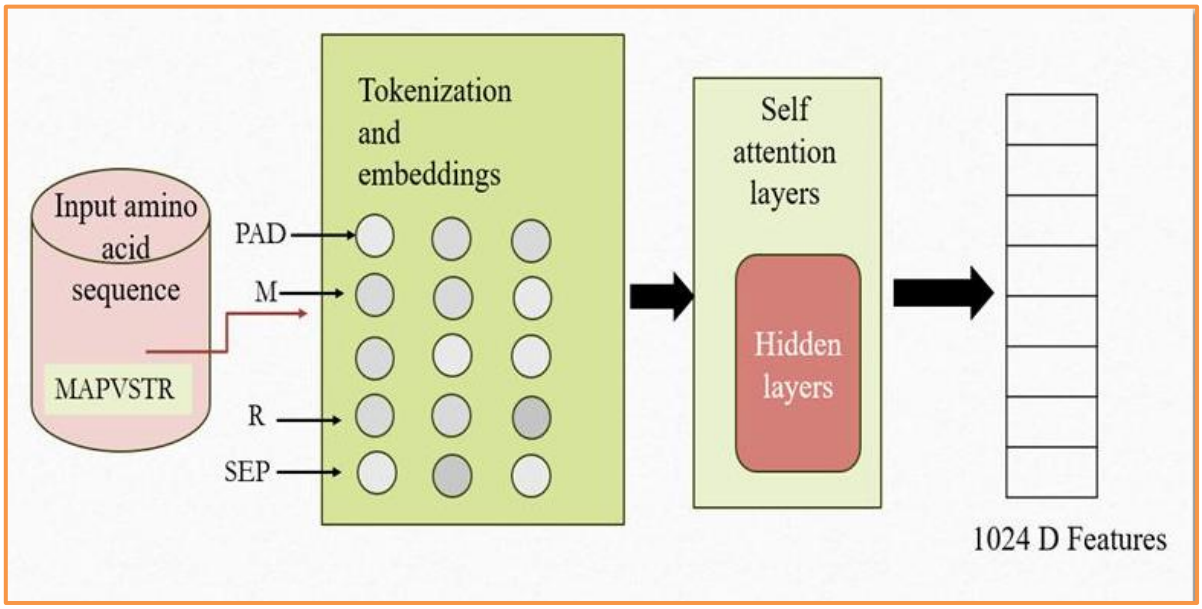

Figure.2: ProtBERT model using word embedding showing how ProtBERT model work.

Transformer Encoder Layers: The embedded sequence is then fed into a deep stack of Transformer encoder layers. Each layer consists of multi-head self-attention mechanism which allow model to target different parts of the sequence simultaneously, capturing complex dependencies between tokens.

After passing through all $\boldsymbol{L}$ encoders, the final output of ProtBERT is the feature matrix $\boldsymbol{X_0}$ which is the refined representation of the input peptide sequence.

Hybrid Input Matrix: The two feature vectors ($[\boldsymbol{F_{ProtBERT}} \parallel \boldsymbol{F_{PsePSSM-DWT}}]$) are concatenated to create the deep neural network's final input, $\boldsymbol{F_{ProtBERT}}$ provides deep contextual representations and $\boldsymbol{F_{PsePSSM-DWT}}$ contributes to extract the evolutionary features.

Feature Selection Using SHAP: Once the hybrid feature vector has been used to train the deep neural network, Deep-ProBind uses SHAP model for the contribution of each feature

associated with model's output. The game-theory method known as SHAP gives each feature a Shapley value that indicates how it affects the prediction of particular input (Fryer *et al.*, 2021). The identification of features from the extracted vector are most important, BorutaSHAP-based wrapper feature selection assesses each feature's contribution to model performance. SHAP selected the top 125 features from a hybrid feature vector with a total dimension of 1244. Features with higher contributions are denoted by red points, and those with lesser contributions are denoted by blue points. The SHAP values are displayed on the horizontal axis; positive values reflect a prediction for protein-binding peptides, whereas negative values point to the class of non-binding peptides.

Model training: To calculate the loss between the true labels and predicted values, the cross-entropy loss function is employed. The loss function is defined as:

$$Loss(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \, log\hat{y}_i) + (1 - y_i) \, log \, (1 - \hat{y}_i)$$

Where:

- n: denotes the batch size of peptide sequences.
- y_i: represents the true label of the peptide sequences
- $\hat{y}_i$: represents the predicted label of the peptide sequences

Deep-ProBind is a classification model, its performance is evaluated using classification metrics, key metrics are used: Accuracy, Precision, Recall, F1 score, Matthew correlation coefficient.

1. $\boldsymbol{Accuracy = \frac{TP+TN}{TP+TN+FP+FN}}$
2. $\boldsymbol{Precision = \frac{TP}{TP+FP}}$
3. $\boldsymbol{Recall = \frac{TP}{TP+FN}}$
4. $\boldsymbol{F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall}}$
5. $\boldsymbol{MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}}$

Where, TP, TN, FP, FN is true positive, true negative, false positive and false negative respectively.

Together, these metrics assess the standard evaluation of the model's ability to accurately classify binding and non-binding peptides.

Evaluation metrics

Table 1.: Evaluation metrics utilized for the evaluation of efficacy of Deep-ProBind.

S. No.	Metrics	Purpose
1.	Accuracy	Quantify the ratio of exact predictions (both positive and negative) among total predictions.
2.	Precision	Quantify the ratio of truly predicted positives among all predicted positives.
3.	Recall (Sensitivity)	Evaluate model's efficiency in recognizing true positives while minimizing false negatives.
4.	F1 Score	Harmonize precision and recall into a unified measure.
5.	Matthews Correlation Coefficient (MCC)	Measure a comprehensive assessment by considering all components of the metrics i.e. TP, TN, FP, FN.

Results and discussion

Sample data containing protein-peptide complex IDs, peptide chain identifiers, UniProt accession numbers, and labelled as protein binding peptide or not as 1,0 respectively was taken. The model performed well with accuracy, Precision, Recall, F1 scores 0.9093, 0.8520, 0.8715,0.8723 respectively. Although Dataset 1 demonstrates a slightly decrement in performance, as compare to the benchmark results reported for Deep-ProBind, the overall

performance remains strong. The MCC value of **0.8866** further confirms that the predictions are reliable and balanced, however in the case of availability of different classes within the dataset.

Table 2.: Values of the accuracy, F1, Precision, Recall, MCC and the sample (Dataset1):

Name	Accuracy	F1 score	Precision	Recall	MCC
Deep-ProBind	0.934	0.934	0.932	0.936	0.87
Dataset1	**0.9093**	**0.8723**	**0.8520**	**0.8715**	**0.8866**

The high accuracy obtained by Deep-ProBind can be attributed by its deep learning algorithm, effectively captures sequence patterns and evolutionary information associated with protein-peptide interactions. The integration of various biological features, deep learning-based algorithm precisely distinguishes binding and non-binding peptides which may not be identified by conventional machine learning based approaches. Overall, these results show that Deep-ProBind is a powerful and efficient computational approach for accurate protein–peptide binding prediction. Its capability to achieve high accuracy along with maintaining performance among various evaluation metrics offers it for large-scale screening of protein-peptide interactions. These approaches can reduce the requirement of extensive experimental validation, thereby saving both time and resources. Consequently, Deep-ProBind play a crucial role in peptide therapeutics, protein engineering, functional annotation, and drug discovery, where accurate identification of protein-binding peptides is essential.

Conclusion

Deep-ProBind represents a significant advancement in protein-peptide interaction prediction. It combines ProtBERT and PsePSSM-DWT to learn complex biological features from peptide sequences, outperforming existing methods and showing strong generalization across datasets. It also utilized SHAP-driven feature selection to enhance interpretability along with precision. However, its training data mainly covers peptides with extreme binding affinities, missing the intermediate range, which could limit broad applicability. Despite these limitations, Deep-ProBind shows great promise in advancing the understanding of protein-peptide recognition mechanisms and has potential applications in therapeutic peptide design, drug discovery, and peptide-based biomarker development.

References

1. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L. and Rost, B. (2021). ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(10): 7112-7127.
2. Fosgerau, K. and Hoffmann, T. (2015). Peptide therapeutics: current status and future directions. *Drug Discovery Today*, **20**(1): 122-128.
3. Fryer, D., Strümke, I. and Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, **9**: 144352-144360.
4. Huang, J., Li, W., Xiao, B., Zhao, C., Zheng, H., Li, Y. and Wang, J. (2024). PepCA: Unveiling protein-peptide interaction sites with a multi-input neural network model. *iScience*, **27**(10).
5. Khan, S., Noor, S., Awan, H. H., Iqbal, S., AlQahtani, S. A., Dilshad, N. and Ahmad, N. (2025). Deep-ProBind: binding protein prediction with transformer-based deep learning model. *BMC Bioinformatics*, **26**(1): 88.
6. Lau, J. L. and Dunn, M. K. (2018). Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & Medicinal Chemistry*, **26**(10): 2700-2707.
7. Marqus, S., Pirogova, E. and Piva, T. J. (2017). Evaluation of the use of therapeutic peptides for cancer treatment. *Journal of Biomedical Science*, **24**(1): 21.

8. Uhlig, T., Kyprianou, T., Martinelli, F. G., Oppici, C. A., Heiligers, D., Hills, D. and Verhaert, P. (2014). The emergence of peptides in the pharmaceutical business: From exploration to exploitation. *EuPA Open Proteomics*, **4**: 58-69.
9. Yin, S., Mi, X. and Shukla, D. (2024). Leveraging machine learning models for peptide–protein interaction prediction. *RSC Chemical Biology*, **5**(5): 401-417.

Acknowledgements

The authors are thankful to CABin Scheme (F. no. Agril. Edn. 4–1/2013-A&P), Indian Council of Agricultural Research, Ministry of Agriculture and Farmers' Welfare, Govt. of India for providing infrastructural support to carry out this research and for creation of Advanced Super Computing Hub for Omics Knowledge in Agriculture (ASHOKA) facility where the work was carried out. The grant of IARI Merit scholarship to PB is duly acknowledged.