ACTA BIOCLINICA

Volumen 15, N° 31. Julio-Diciembre 2025

Artículo Original

Depósito Legal: PPI201102ME3815

Pradeep Kumar Y. y Col.

ISSN: 2244-8136

PREDICTION AND CLASSIFICATION OF NATURAL AND NON-NATURAL RESIDUES IN CELL-PENETRATING PEPTIDES USING A PRE-TRAINED BERT MODEL

Pradeep Kumar Yadalam[1], Soundharya Manogaran[2], Ardila Carlos M.[1,3]

1. Department of Periodontics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai, Tamil Nadu, India.

2. Department of oral biology, Saveetha Dental College and hospitals, Saveetha Dental College, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India.

3. Department of Basic Sciences, Biomedical Stomatology Research Group, Faculty of Dentistry, Universidad de Antioquia, Medellín, Colombia.

Received: 02/27/2025
Accepted: 03/06/2025

EMAIL: martin.ardila@udea.edu.co / pradeepkumar.sdc@saveetha.com

CORRESPONDENCE:
Carlos M. Ardila, Calle 70 No. 52-21, Medellín, Colombia. and Pradeep Kumar Yadalam, Saveetha University, Chennai, Tamil Nadu, India.

ACTA BIOCLINICA

Volumen 15, N° 31. Julio-Diciembre 2025

Artículo Original

Depósito Legal: PPI201102ME3815

Pradeep Kumar Y. y Col.

ISSN: 2244-8136

ABSTRACT

Introduction: Cell-penetrating peptides (CPPs) are amino acids that transport molecular cargo across cellular membranes, making them useful in drug delivery, gene therapy, vaccines, and more. They can cross lipid bilayers, enhance vaccine delivery, and improve immune response. Machine learning can improve drug discovery by predicting CPPs, enhancing drug delivery systems, and personalizing medicine. **Objective:** We aim to predict and classify natural and non-natural residues of cell-penetrating peptides using a pre-trained BERT model. **Methods:** The study used a curated database of over 1,564 validated experimental cell-penetrating peptides (CPPs) with natural and non-natural residues. The datasets were cleaned and extracted using FASTA header identification and regex pattern matching. The extracted sequences were standardized to uppercase and length-based, resulting in 1,547 positive and 286 negative sequences. The study used a numerical vocabulary to tokenize amino acids and tokens and a BERT transformer to convert sequences into dense vectors. The model was trained using a structured looping protocol, including epoch iteration and loss computation. **Results:** The classification model for distinguishing non-native and native residues is evaluated using precision, recall, F1-score, and support metrics. The model strongly understands both classes, minimizing false positives, and has a good trade-off between accuracy and sensitivity. Its overall accuracy is 86%, with consistent performance across both classes. **Conclusion:** An 86% accuracy

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

peptide and protein bioinformatics model distinguishes native and non-native residues. However, it faces limitations like dataset imbalance and overfitting. Future developments will improve data balance, advanced modeling techniques, and biological insights.

KEYWORDS: Natural Language Processing; Peptides; Peptide Mapping; cell-penetrating peptides.

PREDICCIÓN Y CLASIFICACIÓN DE RESIDUOS NATURALES Y NO NATURALES EN PÉPTIDOS PENETRANTES DE CÉLULAS USANDO UN MODELO BERT PRE-ENTRENADO

RESUMEN

Introducción: Los péptidos penetrantes de células (PPCs) son aminoácidos que transportan carga molecular a través de las membranas celulares, haciéndolos útiles en la administración de fármacos, terapia génica, vacunas y más. Pueden cruzar bicapas lipídicas, mejorar la administración de vacunas y mejorar la respuesta inmune. El aprendizaje automático puede mejorar el descubrimiento de fármacos al predecir PPCs, mejorar los sistemas de administración de fármacos y personalizar la medicina. **Objetivo:** Nuestro objetivo es predecir y clasificar residuos naturales y no naturales de péptidos penetrantes de células utilizando un modelo BERT pre-entrenado. **Métodos:** El estudio utilizó una base

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

de datos curada de más de 1.564 péptidos penetrantes de células experimentales validados con residuos naturales y no naturales. Los conjuntos de datos se limpiaron y extrajeron utilizando la identificación de encabezados FASTA y la coincidencia de expresiones regulares. Las secuencias extraídas se estandarizaron a mayúsculas y basadas en la longitud, lo que resultó en 1.547 secuencias positivas y 286 negativas. El estudio utilizó un vocabulario numérico para segmentar aminoácidos y elementos y un transformador BERT para convertir secuencias en vectores densos. El modelo se entrenó utilizando un protocolo de bucle estructurado, que incluye la iteración de épocas y el cálculo de la pérdida. **Resultados:** El modelo de clasificación para distinguir residuos no nativos y nativos se evalúa utilizando métricas de precisión, recuperación, puntuación F1 y soporte. El modelo comprende fuertemente ambas clases, minimizando los falsos positivos, y tiene un buen equilibrio entre precisión y sensibilidad. Su precisión general es del 86%, con un rendimiento consistente en ambas clases. **Conclusión:** Un modelo de bioinformática de péptidos y proteínas con una precisión del 86% distingue los residuos nativos y no nativos. Sin embargo, enfrenta limitaciones como el desequilibrio del conjunto de datos y el sobreajuste. Los desarrollos futuros mejorarán el equilibrio de datos, las técnicas de modelado avanzadas y los conocimientos biológicos.

PALABRAS CLAVE: Procesamiento del Lenguaje Natural; Péptidos; Mapeo de Péptidos; Péptidos Penetrantes de Células.

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

INTRODUCTION

Cell-penetrating peptides (CPPs) are amino acids that facilitate the transport of molecular cargoes across cellular membranes. They are used in drug delivery, gene therapy, vaccines, protein delivery, diagnostics, biotechnology research, neuroscience, antimicrobial agents, cancer therapy, and targeted cell therapies (1). CPPs can cross the lipid bilayer of cells, a barrier for larger molecules like proteins and nucleic acids. They can deliver small molecules, peptides, proteins, or nucleic acids into cells, making them useful in cancer therapies. They can also transport plasmids or mRNA for gene editing, oligonucleotides, siRNAs, and mRNAs for gene silencing and editing technologies (2). CPPs can also enhance the delivery of antigens or adjuvants in vaccines, improve the immune response, and transport functional proteins into cells. However, they must be evaluated for toxicity, intracellular fate, and effectiveness. Despite these challenges, CPPs remain a versatile tool for biological research and oral and periodontal therapeutic applications (3–5).

Biological membranes act as barriers for drugs, making it difficult for new therapies like gene and protein therapy to enter cells. Cell-penetrating peptides (CPPs) offer a promising solution for delivering substances into cells with less toxicity (6,7). CPPs can carry different cargo types and be attached using covalent or non-covalent binding methods. The transport of CPPs across

biological membranes is unclear, but three main pathways have been reported: peptide concentration, peptide sequence, and lipid components in each membrane (8,9). Peptide concentration influences the uptake route of cationic CPPs, with higher concentrations causing rapid cytosolic uptake. The peptide sequence significantly influences cell-penetrating peptide (CPP) activity. Arginine-rich CPPs, including Tat and penetratin, increase local concentrations due to their high positive charge density. Amphipathic CPPs, such as MAPs, require both helical amphipathicity and a length of at least four complete helical turns to efficiently translocate across cell membranes. The positive charge of CPPs is essential for transport, but the charge alone is insufficient. The peptide-to-cell ratio can influence the uptake mode, with higher ratios resulting in direct penetration and endocytosis. Lipid components, such as heparin sulfate proteoglycans or phospholipids, are pivotal in the internalization mechanism (10,11).

CPPs are a potential drug delivery method for treating cancer and diabetes. Machine learning can improve drug discovery by predicting CPPs using various algorithms and datasets 12). CPPs are crucial for drug delivery in cells, and accurately predicting them can save time in experiments. A new model called pLM4CCPs, created using convolutional neural networks, shows improved accuracy and sensitivity compared to

previous models. The best-performing models are ESM-1280 and ProtT5-XL BFD, with pLM4CCPs combining predictions from various models for better peptide classification (13). Predicting the natural residues in CPPs is essential for enhancing drug delivery systems, understanding biomolecular interactions, optimizing synthetic peptides, and personalizing medicine based on variability in effectiveness. BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art model that can predict CPP residues by encoding protein sequences, enabling contextual representation, and facilitating transfer learning. Its capabilities include identifying key residues, extracting features for predictive modeling, handling variability, and multi-task learning. By harnessing BERT to predict CPP residues, researchers can enhance therapeutic strategies, improve drug delivery, and deepen their understanding of the mechanisms of cell penetration, positioning BERT as a pivotal tool in peptide design and evaluation. We aim to predict and classify natural and non-natural residues of cell-penetrating peptides using a pre-trained BERT model.

Materials and Methods

Dataset Retrieval

The peptide structures used in this study were sourced from a curated database, primarily from CPPsite 2.0 (12), which contains over 1,564 validated experimental CPPs with natural residues

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

and 291 peptides with non-natural residues. The analysis involved retrieving two distinct text files, the Positive and Negative Sequences files, using standard Python operations for successful reading.

Data Cleaning and Sequence Extraction

Once the datasets were loaded, comprehensive cleaning procedures were implemented to extract the sequences: FASTA Header Identification: The sequences were parsed by locating FASTA headers, marked by the presence of the " > " symbol. This header precedes sequence data in FASTA format, facilitating the separation of sequence identifiers from their respective sequences.

Extraction Process

Each header and its following sequence data were retrieved through either loop techniques or regex pattern matching. The headers were saved for reference, while sequences underwent cleansing to remove leading and trailing whitespace, improving accuracy.

Standardization of Sequences

The extracted sequences were converted to uppercase for uniformity and reduced case sensitivity issues. Non-standard amino acid filters excluded non-standard amino acids, allowing only the following: A, C, D, E, F, P, H, I, Isoleucine, K, L, M, N, P, Q, R, S, T, V, W, and Y.

Sequence Filtering

The sequences underwent length-based filtering, discarding those shorter than five amino acids for meaningful analysis. The final datasets consisted of 1,547 positive and 286 negative sequences, with the minimum length requirement of five amino acids.

Tokenization and Transformation

Token Insertion: Each sequence was prefixed with a [CLS] token and suffixed with a [SEP] token. This formatting is essential for many transformer-based models, particularly for classification tasks.

Example Transformation

Original Sequence:

"ACDEFGHIKLMNPQRSTVWY"

Tokenized Sequence:

"C ACDEFGHIKLMNPQRSTVWY S"

The model used a numerical vocabulary to map amino acids and tokens, assigning unique indexes to them. Sequences were either padded using a designated token or truncated to a fixed length of 100 tokens. This ensured uniform input sizes and reduced computational complexity. The vocabulary mapping process also created a numerical representation of amino acids.

Model Architecture

The model uses a transformer from BERT, which converts tokenized sequences into dense vector representations. It employs Transformer

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

Encoder layers with multi-head self-attention mechanisms to capture token dependencies. The model's final step is a classifier head that uses the output from the [CLS] token to predict binary classes, specifically identifying whether the peptide is n, using a Multi-Layer Perceptron (MLP).

Hyperparameters

The model was configured with hyperparameters such as vocabulary size, embedding dimension, hidden dimension, number of epochs, learning rate, and batch size. Vocabulary size is the total count of distinct amino acid tokens and special tokens. The embedding dimension is fixed at 64 for effective learning. The hidden dimension is set at 128 for feed-forward layers. Epochs are set at 10 for proper training and generalization. The learning rate is set at 1e-3.

Training Loop

The model was trained using a structured looping protocol, including epoch iteration and loss computation. The training set included positive and negative sequences for each epoch. Binary cross-entropy loss was calculated for both sets, collected across batches, and averaged for performance measurement.

Results

The classification model for distinguishing between non-native and native residues is evaluated using precision, recall, F1-score, and support metrics. The model's precision is around

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

0.86, indicating a strong understanding of both classes and minimizing false positives. The recall is 86% for non-native and 86% for native, indicating a strong sensitivity to both classes. The F1-score balances precision and recall, indicating a good trade-off between identifying true positives and minimizing false positives. The model's support is 150 instances of non-native residues and 217 instances of native residues, despite an imbalance in support. The overall accuracy is 86%, indicating the model's ability to generalize well to unseen data and categorize both classes effectively. The macro and weighted averages for precision, recall, and F1-score are all 0.86, indicating consistent performance across both classes. The model's robust and balanced performance suggests its reliability in biological applications, particularly in peptide classification and bioinformatics.

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

Table 1 shows the accuracy of the transformer-based BERT model.

Class	Precision	Recall	F1-Score	Support
Non-native (0)	0.87	0.86	0.86	150
Native (1)	0.85	0.86	0.85	217
Accuracy	------	-----	0.86	367
Macro Avg	0.86	0.86	0.86	367
Weighted Avg	0.86	0.86	0.86	367

Data Processing

The initial step involved cleaning and filtering sequences extracted from two separate files, resulting in approximately 1,547 positive and 286 negative sequences. The following steps outline the data processing and model development pipeline:

1. Tokenization:

-Subsequently, tokenization of these sequences was performed. Special tokens, specifically [CLS] and [SEP], were integrated into the sequences, which were then padded or truncated to a uniform length of 100 tokens.

-This step converted the sequences into numerical indices, utilizing a pre-defined vocabulary

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

2. Model Creation:

-A transformer-based classifier was constructed, which incorporated an embedding layer, transformer encoders, and a multi-layer perceptron (MLP) classifier

3. Training:

-The model underwent a training process spanning 10 epochs.

-Throughout this training, both training and validation losses were meticulously recorded. The losses declined from approximately 0.45 to 0.34 across both sets, indicating improved model performance. The best-performing model was preserved at the point of minimal validation loss

4. Evaluation:

-Upon completion of the training phase, a comprehensive classification report was generated, revealing an impressive overall accuracy of around 86%.

-Critical to this analysis was confirming balanced performance between natural and non-natural residues with classes, underscoring the classifier's effectiveness and reliability. Overall, the methodology employed demonstrates a rigorous approach to data processing, classification, and evaluation, leading to significant achievements in model performance.

Figure 1 shows how the training and validation loss decreased over the 10 training epochs. We can observe that training and validation loss steadily decreased throughout the training process. The final training loss was approximately 0.32, while the validation loss was 0.33. There is no significant overfitting, as the validation loss closely follows the training loss.

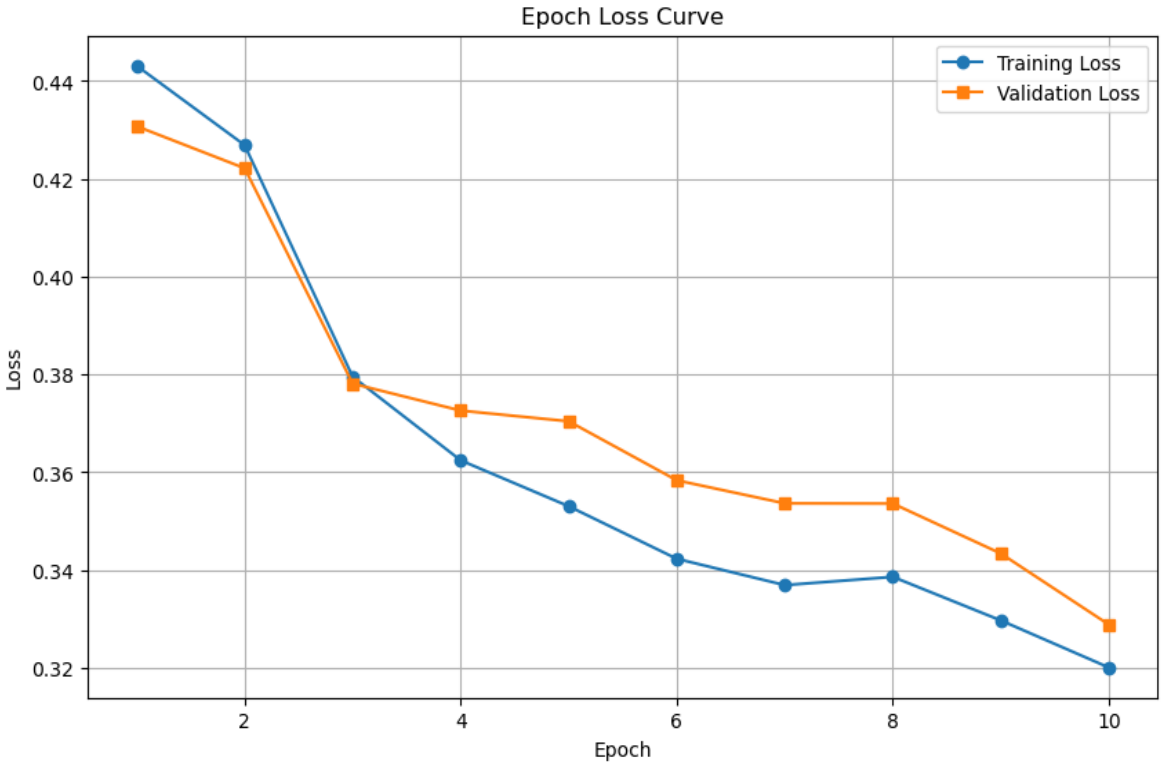

Figure 1. Epoch Loss Curve: Training versus Validation

Figure 2 shows the ROC curve of the model.

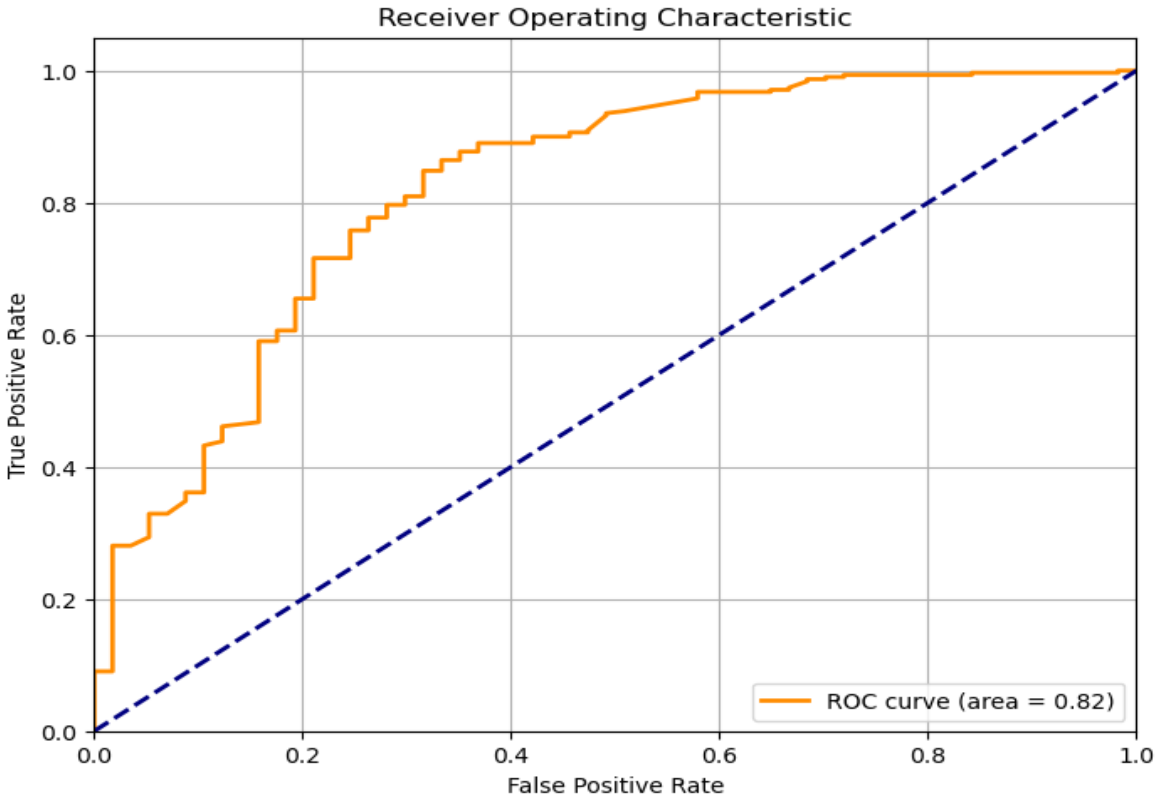

Figure 2. Receiver Operating Characteristic (ROC) Curve

The ROC (Receiver Operating Characteristic) curve plots the True Positive Rate against the False Positive Rate at various threshold settings. The AUC (Area Under Curve) of 0.92 indicates excellent classification performance (a perfect classifier would have an AUC of 1.0).

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

Discussion

The delivery of functional proteins to cells presents a promising therapeutic strategy, particularly for diseases linked to protein dysfunction. Unlike small-molecule drugs, which can have adverse effects and often fail to replicate the specific roles of proteins, protein therapeutics offer a safer alternative without requiring genome modifications (14,15). Additionally, they tend to have shorter development timelines and broader patent protection. However, proteins' large and hydrophilic nature limits their direct cellular uptake. CPPs have emerged as effective tools for facilitating the delivery of these proteins into cells, utilizing mechanisms such as endocytosis. Recent advancements, particularly with cyclic CPPs, have shown improved efficiency in delivering proteins and RNA, indicating that modifications in their design can further enhance their delivery capabilities (16).

Bert-based classification model for distinguishing between non-native and native residues is evaluated using precision, recall, F1-score, and support metrics. The model has a precision of 0.86, indicating strong understanding and minimizing false positives. It recalls 86% for non-native and 86% for native, indicating sensitivity to both classes. The F1-score balances precision and recall, indicating a good trade-off between identifying true positives and minimizing false positives. The model's overall accuracy is 86%, indicating its reliability. The macro and weighted averages for precision, recall, and F1-score are 0.86,

similar to this study exploring machine learning models like SVM, Random Forest, J48, naïve Bayes, and SMO for analyzing atom composition and chemical descriptors. The Random Forest model achieved the highest accuracy with 92.33% and an AUROC of 0.98 on a validation dataset (17). The CPP1708 dataset is the largest reliable database of CPPs to date, with GraphCPP demonstrating superior predictive performance compared to previous methods. The model achieved a 92.8% and 23.3% improvement in Matthews correlation coefficient and AUC measures compared to the next best model. GraphCPP's ability to learn peptide representations was demonstrated through t-distributed stochastic neighbor embedding plots (18,19). It maintains high confidence in predictions for peptides shorter than 40 amino acids. Also similar to one more study showed that his study proposes a feature fusion-based prediction model, where the protein pre-trained language models ProtBERT and ESM-2 are used as feature extractors, and the extracted features from both are fused to obtain a more comprehensive and effective feature representation, which is then predicted by linear mapping. Validated by many experiments on public datasets, the method has an AUC value as high as 0.983 and shows high accuracy and reliability in cell-penetrating peptide prediction (20).

The current classification model for distinguishing between native and non-native residues has strong predictive performance, but there are several areas

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

for improvement. These include enhancing feature engineering, exploring deep learning techniques, implementing data augmentation techniques, conducting cross-validation and robustness testing, applying transfer learning from larger datasets, comparing the model against other state-of-the-art classification models, integrating with biological experiments, assessing beyond binary classification, and exploring interpretability frameworks (1,21). However, the model has limitations such as an imbalanced dataset, potential overfitting, feature limitations, bias in data, biological context, scalability, and limited exploration of biological relevance. The inherent class imbalance may affect the model's ability to generalize to underrepresented classes, and continuous validation through independent datasets is critical. The model's efficiency may be limited if features do not capture relevant biological signals. Any bias in the dataset, such as the selection of residues or environmental conditions not represented in the training data, could adversely affect the model's predictions.

The model does not account for potential biological complexities, such as the influence of post-translational modifications or protein folding, which could impact residue functionality beyond the classification outcome (22). The current architecture may not scale efficiently for larger datasets without optimization, particularly if future studies involve extensive peptide libraries or high-throughput data. Lastly, the focus on

Volumen 15, N° 31. Julio-Diciembre 2025
Depósito Legal: PPI201102ME3815
ISSN: 2244-8136

classification performance may overlook the biological relevance or function of the classified residues, necessitating a deeper analysis of implications in biological systems. Addressing these limitations will be crucial for establishing a more robust and reliable tool for biological applications and furthering research in peptide classification and bioinformatics.

Conclusion

A classification model with an accuracy of 86% distinguishes between native and non-native residues in peptide and protein bioinformatics. However, the model faces limitations such as dataset imbalance, overfitting, and a lack of comprehensive biological context. Future developments will focus on improving data balance, exploring advanced modeling techniques, and integrating biological insights to enhance the model's applicability and accuracy. This will create a more robust tool that enhances our understanding of protein functionality in complex biological systems. This work lays the groundwork for further explorations and collaborations in bioinformatics, bridging the gap between computational predictions and experimental validations.

Conflict of Interest

The authors have no conflicts of interest to declare

Authors contribution

Conceptualization: Pradeep Kumar Yadalam, Soundharya Manogaran and Carlos M. Ardila. Methodology: Pradeep Kumar

Yadalam, Soundharya Manogaran and Carlos M. Ardila. Software: Pradeep Kumar Yadalam. Formal analysis: Pradeep Kumar Yadalam, Soundharya Manogaran and Carlos M. Ardila. Investigation: Pradeep Kumar Yadalam, Soundharya Manogaran and Carlos M. Ardila. Data curation: Pradeep Kumar Yadalam, Soundharya Manogaran and Carlos M. Ardila. Writing-original draft preparation, Pradeep Kumar Yadalam, Soundharya Manogaran and Carlos M. Ardila; writing-review and editing, Pradeep Kumar Yadalam and Carlos M. Ardila; administration: Pradeep Kumar Yadalam

Ethics approval: Not applicable

REFERENCES

1. Derakhshankhah H, Jafari S. Cell penetrating peptides: A concise review with emphasis on biomedical applications. Biomed Pharmacother. 2018;108:1090-1096.

2. Moreno-Vargas LM, Prada-Gracia D. Exploring the Chemical Features and Biomedical Relevance of Cell-Penetrating Peptides. Int J Mol Sci. 2024;26(1):59.

3. Ramamurthy J, Nedumaran N. Evaluation of C-reactive Protein and Interleukin-6 Levels in Periodontitis Patients: An Experimental Study. World J Dent. 2025;15 (9):772–6.

4. Burra Anand D, Ramamurthy J, Kannan B, Jayaseelan VP, Arumugam P. N6-methyladenosine-mediated overexpression of TREM-1 is associated with periodontal disease. Odontology. 2024 Sep 26. doi: 10.1007/s10266-024-01009-w.

5. Ramamurthy J. Evaluation of Antimicrobial Activity of Nanoformulated Grape Seed Oil against Oral Microbes: An In Vitro Study. World J Dentt. 2024;15:44–7.

6. Kumar V, Agrawal P, Kumar R, Bhalla S, Usmani SS, Varshney GC, Raghava GPS. Prediction of Cell-Penetrating Potential of Modified Peptides Containing Natural and Chemically Modified Residues. Front Microbiol. 2018;9:725.

7. Patel SG, Sayers EJ, He L, Narayan R, Williams TL, Mills EM, et al. Cell-penetrating peptide sequence and modification dependent uptake and subcellular distribution of green florescent protein in different cell lines. Sci Rep. 2019;9(1):6298.

8. Du JJ, Zhang RY, Jiang S, Xiao S, Liu Y, Niu Y, et al. Applications of cell penetrating peptide-based drug delivery system in immunotherapy. Front Immunol. 2025;16:1540192.

9. Miwa A, Kamiya K. Cell-Penetrating Peptide-Mediated Biomolecule Transportation in Artificial Lipid Vesicles and Living Cells. Molecules. 2024;29(14): 3339

10. Gong X, Han Y, Wang T, Song G, Chen H, Tang H, et al. Cell-Penetrating Peptide Induced Superstructures Triggering Highly Efficient Antibacterial Activity. Adv Mater. 2025;37(4):e2414357.

11. Sutcliffe R, Doherty CPA, Morgan HP, Dunne NJ, McCarthy HO. Strategies for the design of biomimetic cell-penetrating peptides using AI-driven in silico tools for drug delivery. Biomater Adv. 2025;169:214153.

12. Manavalan B, Subramaniyam S, Shin TH, Kim MO, Lee G. Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. J Proteome Res. 2018;17(8):2715–26.

13. Zhu L, Chen Z, Yang S. EnDM-CPP: A Multi-view Explainable Framework Based on Deep Learning and Machine Learning for Identifying Cell-Penetrating Peptides with Transformers and Analyzing Sequence Information. Interdiscip Sci. 2024. doi: 10.1007/s12539-024-00673-4.

14. Zhang H, Zhang Y, Zhang C, Yu H, Ma Y, Li Z, et al. Recent Advances of Cell-Penetrating Peptides and Their Application as Vectors for Delivery of Peptide and Protein-Based Cargo Molecules. Pharmaceutics. 2023;15(8):2093

15. Hardan L, Chedid JCA, Bourgi R, Cuevas-Suárez CE, Lukomska-Szymanska M, Tosco V, et al. Peptides in Dentistry: A Scoping Review. Bioengineering. 2023;10(2): 214

16. Bermúdez M, Hoz L, Montoya G, Nidome M, Pérez-Soria A, Romo E, et al. Bioactive Synthetic Peptides for Oral Tissues Regeneration. Front Mater. 2021;8:655495.

17. Ramasundaram M, Sohn H, Madhavan T. A bird's-eye view of the biological mechanism and machine learning prediction approaches for cell-penetrating peptides. Front Artif Intell. 2024;7:1497307.

18. Kumar N, Du Z, Li Y. pLM4CPPs: Protein Language Model-Based Predictor for Cell Penetrating Peptides. J Chem Inf Model. 2025;65(3):1128–39.

19. Imre A, Balogh B, Mándity I. GraphCPP: The new state-of-the-art method for cell-penetrating peptide prediction via graph neural networks. Br J Pharmacol. 2025;182(3):495–509.

20. Zhang F, Li J, Wen Z, Fang C. FusPB-ESM2: Fusion model of ProtBERT and ESM-2 for cell-penetrating peptide prediction. Comput Biol Chem. 2024;111:108098.

21. Gu ZF, Hao YD, Wang TY, Cai PL, Zhang Y, Deng KJ, et al. Prediction of blood-brain barrier penetrating peptides based on data augmentation with Augur. BMC Biol. 2024;22(1):86.

22. Kumar A, Chadha S, Sharma M, Kumar M. Deciphering optimal molecular determinants of non-hemolytic, cell-penetrating antimicrobial peptides through bioinformatics and Random Forest. Brief Bioinform. 2024;26(1):bbaf049