JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

pLM4CPPs: Protein Language Model-Based Predictor for Cell Penetrating Peptides

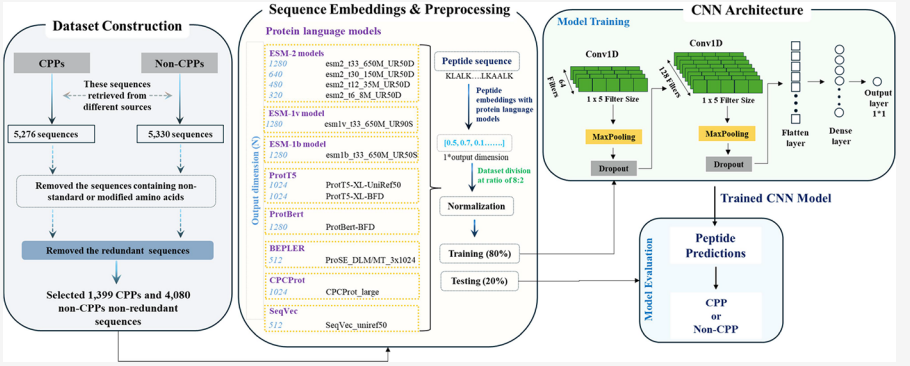
Published as part of Journal of Chemical Information and Modeling special issue "Harnessing the Power of Large Language Model-Based Chatbots for Scientific Discovery".

Nandan Kumar, Zhenjiao Du, and Yonghui Li*

Cite This: https://doi.org/10.1021/acs.jcim.4c01338 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via KANSAS STATE UNIV on January 29, 2025 at 13:59:32 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.



ABSTRACT: Cell-penetrating peptides (CPPs) are short peptides capable of penetrating cell membranes, making them valuable for drug delivery and intracellular targeting. Accurate prediction of CPPs can streamline experimental validation in the lab. This study aims to assess pretrained protein language models (pLMs) for their effectiveness in representing CPPs and develop a reliable model for CPP classification. We evaluated peptide embeddings generated from BEPLER, CPCProt, SeqVec, various ESM variants (ESM, ESM-2 with expanded feature set, ESM-1b, and ESM-1v), ProtT5-XL UniRef50, ProtT5-XL BFD, and ProtBERT. We developed pLM4CCPs, a novel deep learning architecture using convolutional neural networks (CNNs) as the classifier for binary classification of CPPs. pLM4CCPs demonstrated superior performance over existing state-of-the-art CPP prediction models, achieving improvements in accuracy (ACC) by 4.9−5.5%, Matthews correlation coefficient (MCC) by 9.3−10.2%, and sensitivity (Sn) by 14.1−19.6%. Among all the tested models, ESM-1280 and ProtT5-XL BFD demonstrated the highest overall performance on the kelm data set. ESM-1280 achieved an ACC of 0.896, an MCC of 0.796, a Sn of 0.844, and a specificity (Sp) of 0.978. ProtT5-XL BFD exhibited superior performance with an ACC of 0.901, an MCC of 0.802, an Sn of 0.885, and an Sp of 0.917. pLM4CCPs combine predictions from multiple models to provide a consensus on whether a given peptide sequence is classified as a CPP or non-CPP. This approach will enhance prediction reliability by leveraging the strengths of each individual model. A user-friendly web server for bioactivity predictions, along with data sets, is available at https://ry2acnp6ep.us-east-1.awsapprunner.com. The source code and protocol for adapting pLM4CPPs can be accessed on GitHub at https://github.com/drkumarnandan/pLM4CPPs. This platform aims to advance CPP prediction and peptide functionality modeling, aiding researchers in exploring peptide functionality effectively.

1. INTRODUCTION

The delivery of various cargoes (including small molecules, oligonucleotides, proteins, etc.) into cells has the potential to revolutionize future therapeutics.[1,2] Researchers have made significant progress in developing new methods to deliver therapeutic compounds across the cell membrane.[3−5] Antimicrobial peptides (AMPs) are a naturally occurring class of molecules known for their microbicidal properties. They are often positively charged (cationic) and have a mixed composition of water-loving (hydrophilic) and water-fearing (hydrophobic) regions.[6,7] Due to these properties, some AMPs have emerged as a promising subgroup within a larger category called cell-penetrating peptides (CPPs).[8] In general, they are short peptides, typically containing 5−50 amino acids. They

Received: July 28, 2024
Revised: January 14, 2025
Accepted: January 15, 2025

© XXXX American Chemical Society

possess a unique ability to transport various molecules, including small molecules, proteins, and even large particles, directly into cells with minimal harm to the cell membrane.[9−11] In recent years, CPPs have emerged as promising drug delivery vehicles, enabling the transport of pharmacologically active molecules such as oligonucleotides,[12] plasmid DNA,[13] short interfering RNA,[14] peptide nucleic acid,[15] peptides,[16,17] proteins,[18] and nanoparticles,[19] across the membrane. The number of known CPP sequences has increased rapidly, with new modified CPPs being developed to improve their stability and bioavailability. These novel CPPs are most often derived from the existing proteins and further optimized to be the shortest peptides having maximum transportation capability across the cell membrane.[10,20] Advances in proteomic technologies, including next-generation sequencing, have significantly enhanced the understanding of CPPs.[21] Researchers leverage gene editing and phage display to design new CPP candidates[22] and the techniques like fluorescent labeling and Caco-2 cell arrays have been used then to assess their membrane penetration ability.[23] However, traditional in vitro assays for identifying optimal CPPs remain slow and labor-intensive.[24] Moreover, CPP effectiveness is linked to both their sequence and physical/chemical properties.[10] The relationship among sequence, properties, and effectiveness makes computational methods a powerful approach for identifying promising CPP candidates. This approach significantly reduces the experimental burden on researchers.[25]

Recently, machine learning (ML) approaches have gained significant interest in predicting the structure and function of peptides, protein, and other biomolecules.[26] These methods offer fast, reliable, and accurate predictions based solely on the sequences, without requiring additional information.[27] Several ML models such as CPPpred,[28] CellPPD,[29] C2Pred,[30] SkipCPP-Pred,[31] CPPred-RF,[32] MLCPP-2.0,[25] and BChemRF-CPPred[33] have been reported for predicting CPPs. These models utilize diverse methods, leveraging features such as sequence composition, physicochemical properties, dipeptide composition, motif information, and biochemical features. Additionally, techniques such as minimum Redundancy Maximum Relevance (mRMR), incremental feature selection (IFS), and analysis of variance are employed to refine these features and improve model performance. This variety of approaches highlights the flexibility of ML in CPP prediction and the ongoing quest for enhanced accuracy and reliability. While existing methods provide a comprehensive overview of algorithms, feature encodings, and evaluation strategies,[34] their reliance on a limited set of features remains a key constraint. This restricted representation hinders their ability to achieve optimal prediction accuracy. The advent of transformers and large language models has introduced nonhandcrafted (self-supervised) features that can outperform traditional handcrafted features.[35] This shift indicates that relying solely on a limited set of hand-crafted features may not be sufficient for optimal classification performance, despite the extensive use of unsupervised learning in CNNs before the rise of transformers. Advancements in natural language processing (NLP) have spurred the development of protein language models (PLMs) for downstream protein sequence tasks. These models, known for their ability to capture complex patterns in data, are being adapted for bioinformatics tasks. Inspired by the success of transformer-based models in NLP tasks, researchers have explored the use of pretrained PLMs for various protein/peptide sequence analysis and prediction tasks.[36−43] For instance, a study by Martnez-Mauricio et al. focuses on classifying AMPs using embeddings derived from ESM-2 models and highlights the effectiveness of different embedding dimensions (640- and 1280-dimensional) from ESM-2 models. These embeddings were found to yield statistically better performances in quantitative structure−activity relationship (QSAR) models compared to other methods.[44] Furthermore, recent advancements have introduced graph-based deep learning approaches and PLMs for AMP classification, which can be further explored for predicting CPPs. García-Jacas and co-workers demonstrated the effectiveness of graph-based models with ESM-2 features for AMPs, which suggests potential for similar approaches in CPP prediction.[45]

In this study, we leverage a diverse set of protein embedding techniques to capture comprehensive and informative representations from peptide sequences. These techniques include BEPLER embedding for capturing structural information from protein sequences using a multitask learning framework,[46] CPCProt embedding for leveraging contrastive predictive coding to maximize mutual information between local and global sequential embeddings,[47] SeqVec embedding for biophysical properties using the ELMo model from natural language processing,[48] ESM-based embeddings (ESM-2, ESM-1b, ESM-1v), which utilize BERT-based architectures for complex relationship learning,[41,49,50] and ProtT5-based embeddings (ProtT5-XL-UniRef50, ProtT5-XL-BFD, ProtT5-Port-BERT), which employ transformer architectures adapted from powerful NLP models for protein sequence modeling.[40,42,51] To the best of our knowledge, existing studies have not employed PLMs for the feature representation of CPP sequences for developing prediction models. This work aims to address this gap by comparing different PLMs and developing PLM-based models in conjunction with CNN for CPP prediction. This multifaceted approach has the capability to capture a wider range of sequence features relevant to peptide function identification, potentially leading to more robust models that overcome limitations observed in existing methods.

2. MATERIALS AND METHODS

2.1. Data Set Construction. Due to the disparate nature of the training data sets used by the existing methods, we constructed a comprehensive data set from several well-established resources like CPPsite2.0, C2Pred, CellPPD, MLCPP 2.0, and KELM-CPPpred. Positive examples (CPPs) were predominantly experimentally validated sequences. Negative examples (non-CPPs) were sourced from the same databases and included sequences identified through experimental validation, computational predictions, or literature references indicating that they lack cell-penetrating properties.[25,29,30,52,53] A total of 10,606 sequences were collected. These sources were chosen due to their comprehensive collections of CPP and non-CPP sequences. First, the positive and negative samples (CPPs and non-CPPs) were grouped, resulting in 5276 CPPs and 5330 non-CPPs. Sequences containing specific modified peptide sequences were identified and subsequently removed to maintain consistency and compatibility with standard sequence-based analysis methods. For compatibility with protein embedding techniques and machine learning models, we excluded peptide sequences containing nonstandard or modified residues. We further performed a redundancy check to ensure nonredundant

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

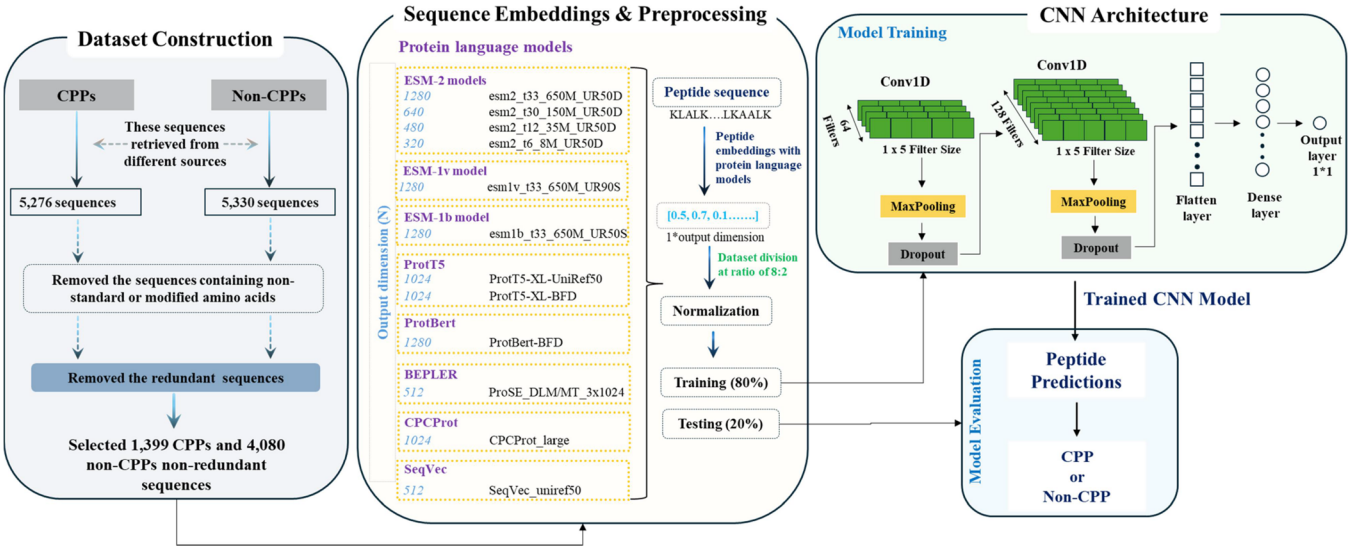


Figure 1. Schematic framework for predicting peptides as CPPs and non-CPPs by integrating protein language models (pLMs) and convolutional neural networks (CNNs). Peptides of any length are encoded into varied dimensional embeddings by pLMs and then fed into the CNN model. The first CNN layer consists of 64 filters, each undergoing 1D convolution with a kernel size of 5 and a ReLU activation function. This is followed by down sampling via a max-pooling layer, resulting in a 64 × 640 feature matrix. The next convolutional layer has 128 filters followed by another max-pooling layer, resulting in a 128 × 320 feature matrix. This feature matrix is flattened into a 1D vector and passed into a dense layer with 256 neurons and ReLU activation. The final output layer consists of one neuron with a sigmoid function for predicting whether the peptide is a CPP or a non-CPP.

sequences for further analysis. This involved using in-house-developed Python scripts that efficiently identified and removed duplicate sequences. Initially, redundant sequences within the CPP and non-CPP groups were removed. Subsequently, a second check was conducted to eliminate redundancies between the CPP and non-CPP sequences, ensuring the quality of the data set for optimal machine learning performance. Following these steps, the nonredundant CPP and non-CPP sequences were labeled as 1 and 0, respectively. This resulted in a final data set comprising 1399 CPPs and 4080 non-CPPs, ready for machine learning model development. To compare existing methods with the method developed in this study, benchmark validation data sets were used. These data sets were downloaded from independent data sets provided in Pandey's work[53] and were referred to as "kelm" for convenience.

2.2. Protein Language Models for Embedding the CPP Sequences and Analysis. This study comparatively explores the application of a diverse array of advanced PLMs to generate embeddings in the context of predicting CPPs. The models selected encompassed various architectures and output vector dimensions (feature dimensions) to capture distinct aspects of the peptide sequences. Among the models employed was BEPLER, which utilizes a transformer-based approach to generate 121-dimensional embeddings. CPCProt, on the other hand, utilize 512-dimensional embeddings. These models encode biological sequence data into high-dimensional numerical representations using deep learning techniques. Additionally, SeqVec, known for its bidirectional LSTM architecture, was used to generate 1024-dimension embeddings, capturing both short- and long-range dependencies within sequences that are crucial for the prediction. ProtT5-XL, trained on comprehensive data sets like UniRef50 and Big Fantastic Database (BFD), adopted a Transformer-based architecture to yield embeddings with 1024 dimensions, emphasizing scalability and performance in handling large-scale protein sequence data. A variant, ProtT5-Port-BERT, incorporated a BERT-style architecture to further enhance portability and efficiency in generating protein sequence embeddings.

This research also investigated variations within the Evolutionary Scale Modeling (ESM) family, specifically, ESM-1b and ESM-1v. These models excel at generating protein sequence embeddings with a dimensionality of 1280. They achieve this by incorporating evolutionary information directly into their architecture. This unique approach strengthens their ability to capture complex sequence patterns crucial to predicting peptides. Furthermore, the investigation encompassed features with evolutionary information using the family of ESM-2 models. This family of models includes six pretrained models with 6, 12, 30, 33, 36, and 48 layers, respectively. These models scale up to 8 million, 35 million, 150 million, 650 million, 3 billion, and 15 billion parameters, respectively. The training data comprised 65 million unique sequences,[41] more than double the amount used to pretrain the previous ESM-1b model (27.1 million unique protein sequences).[50] The implementations of BEPLER,[46] CPCProt,[47] and SeqVec[48] embeddings were sourced from the Bio Embeddings Python library.[54] For the ESM2 variants with dimensions of 320, 480, 640, and 1280, we leveraged the publicly available code and pretrained models hosted on the ESM GitHub repository (https://github.com/facebookresearch/esm).[42] ProtT5-XL with UniRef50 and BFD, along with ProtBERT, were sourced from the ProtTrans GitHub repository (https://github.com/agemagician/ProtTrans).[40] The output of the embedding models is a matrix N x M per peptide sequence, where N is the number of peptides and M is the embedding size (i.e., the dimensionality of the vector capturing the features of the peptide).

While larger PLMs with higher feature dimensions have the potential to extract more intricate information from sequences, improving performance in downstream tasks,[39,41,42] studies by

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Martínez-Mauricio et al. have demonstrated that higher dimensions do not always result in better model performance.[44] This highlights the need to manage the "curse of dimensionality", particularly in smaller data sets, where higher feature dimensions can hinder performance. However, in the case of allergen prediction, larger PLMs with higher feature dimensions have demonstrated superior performance.[39] Specifically, ESM-2 with a dimension of 2560 achieved the best performance, illustrating that in larger data sets with rich feature spaces, high-dimensional embeddings can effectively improve model accuracy. This underlines the importance of balancing feature dimension and data set size for optimal model performance. This study is the first of its kind to utilize various PLMs with different dimensions for representing CPP and non-CPP sequences and building deep learning models. The schematic framework of pLM4CPPs integrating pLMs and CNNs for predicting CPPs is shown in Figure 1. To assess the effectiveness of considered PLMs in representing CPPs for bioactivity prediction, Uniform Manifold Approximation and Projection (UMAP) was used to visualize the high-dimensional embeddings in a two-dimensional space.[55] The sequence logo was generated using the positive peptides from the kelm dataset and the active peptides from the dataset used for training and testing the model, using the MEME Suite.[56] These sequences were aligned, and the frequency of each amino acid at every position was calculated. The sequence logo was then constructed to visualize these frequencies, with the height of each letter representing the relative frequency of the corresponding amino acid at that position.

2.3. Data Preparation for Model Training and Evaluation. Following the acquisition of protein sequence embeddings, the data set was split into training and testing sets using an 80:20 ratio, ensuring random distribution with a fixed random seed for reproducibility. Specifically, 80% of the total CPP sequences (1119 sequences) and 80% of the total non-CPP sequences (3264 sequences) were allocated to the training set, while the remaining 20% (280 CPP sequences and 816 non-CPP sequences) were designated for the test set. To ensure an independent evaluation of the model's performance on unseen data, the kelm data set used for model validation was kept separate and not included in either the training or test data sets. To ensure that the feature values were on a similar scale, we normalized the data using the MinMaxScaler from the scikit-learn library. The scaler was fitted to the training data and then applied to both the training and testing sets. This preprocessing step helped improve the performance and convergence of the machine learning models used in subsequent analysis. This normalization ensured consistency in the feature space between training and testing, allowing for a fair evaluation of the model performance. Subsequently, we rigorously compared the effectiveness of different embeddings in conjunction with CNNs for CPP prediction and development of pLM4CPPs.

2.4. Convolutional Neural Network Architecture. The CNN model architecture is designed to process the peptide sequence embeddings generated from the considered PLMs. The architecture consisted of stacked layers that progressively extracted increasingly complex features from the input embeddings. The first layer was a one-dimensional convolution with 64 filters, stride 1 and a kernel size of 5, designed to capture local patterns within the embeddings. Batch normalization was used to stabilize and accelerate training by normalizing the activations of the previous layer (Figure 1). A ReLU activation function was then applied to introduce nonlinearity. Subsequently, a max-pooling layer with a size of 2 reduced the dimensionality, focusing on the most salient features. To further enhance the network's ability to learn intricate patterns, a dropout rate of 0.25 was implemented after each max-pooling layer to prevent overfitting and improve generalization. This process was repeated with a second one-dimensional convolutional layer comprising 128 filters and the same kernel size of 5, followed by batch normalization, ReLU activation, and max-pooling. The flattened output from the convolutional layers was then fed into fully connected layers. A dense layer with 256 neurons and ReLU activation was incorporated to learn higher-level representations. Dropout regularization with a rate of 0.5 was applied before the final dense layer with a sigmoid activation function, which yielded the predicted probability of a peptide sequence being a CPP. The CNN architecture was optimized by using the Adam optimizer with a learning rate of 0.001, configured to minimize binary cross-entropy loss during training. Learning rate scheduling was implemented using a step decay function, halving the learning rate every 10 epochs to facilitate convergence. Early stopping based on validation accuracy with a patience of 20 epochs was employed to prevent overfitting. Model checkpointing ensured that only the best-performing model based on validation accuracy was saved for subsequent evaluation. Finally, class weights were adjusted to address class imbalance by assigning a higher weight to CPPs, improving the model sensitivity for this class.

2.5. Model Evaluation. To assess the performance of the model, we first predicted the probabilities for each sample and optimized the threshold using the Matthews correlation coefficient (MCC). This optimized threshold helped to convert probabilities into binary predictions. We then calculated several key metrics. Accuracy (ACC) was determined as the proportion of true results (both true positives (TP) and true negatives (TN)) among all cases examined. Sensitivity (Sn), or recall, measured the proportion of actual positives that were correctly identified. Specificity (Sp) was calculated as the ratio of true negatives (TN) to the sum of true negatives and false positives (FP), indicating that the proportion of actual negatives was correctly identified. The MCC assessed the quality of binary classifications by considering true positives, false positives, false negatives, and false negatives (FN). The area under the receiver operating characteristic curve (AUC) represented the model's ability to distinguish between classes. Finally, Balanced Accuracy (BACC), defined as the average of sensitivity and specificity, provided a balanced performance measure, particularly valuable for imbalanced data sets. These metrics collectively provided a comprehensive assessment of the predictive capabilities of the model on the test and external kelm data set.

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

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{BACC} = 0.5 \times \mathrm{Sn} + 0.5 \times \mathrm{Sp}$$

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} \times \mathrm{FN})}}$$

2.6. Web Server Deployment and Resources. A user-friendly web server for predicting CPPs is available at https://ry2acnp6ep.us-east-1.awsapprunner.com. This server leverages Amazon Web Services (AWS) App Runner and the Elastic Container Registry (ECR) for deployment. The web interface was developed using HTML and CSS, while Flask (version 2.2.2) facilitated model deployment. Due to computational resource limitations, only the top-performing CPP prediction models (excluding the largest model) are currently available on the server. Users can combine predictions from these models to achieve a consensus classification, increasing the confidence of their CPP predictions. The server accepts data uploads in various formats (XLS, XLSX, FASTA, and TXT) for convenient large-scale processing. For comprehensive development resources, including embedding generation, model training, evaluation scripts, and protocols for predicting new peptide activity using the pLM4CCPs models, please refer to the GitHub repository: https://github.com/drkumarnandan/pLM4CPPs.

3. RESULTS AND DISCUSSION

3.1. Peptide Sequence Analysis. We analyzed the sequences of CPPs and non-CPPs in our data set to understand their length distribution and amino acid composition (Figures 2, S1, and S2). This analysis is crucial because the effectiveness of machine learning models heavily relies on the quality of their training data. Previous studies suggested an enrichment of arginine (R) in CPPs compared to non-CPPs, with lysine (K) and leucine (L) also showing significant differences.[20,32] We investigated these observations in our comprehensive data set to explore the potential length and amino acid preferences that differentiate CPPs and non-CPPs. The length distribution of CPPs in our data set displayed a notable preference for sequences between 10 and 20 amino acids, with peaks observed at 10−15 and 15−20 residues, followed by 5−10 and 20−25 residues (Figure 2A). This distribution suggests that CPPs might optimize their cellular uptake mechanisms within these specific length constraints. Conversely, non-CPP sequences exhibited a broader range of lengths, indicating a less defined length preference (Figure 2A). The experimentally validated kelm data set also displayed a similar length distribution (Figure S2A). The raw count of amino acid residues in CPPs and non-CPPs (Figure S1) could be biased due to differences in the sequence length and number. To address this, we calculated the percentage distribution, which normalizes the composition relative to the total number of analyzed residues for CPPs and non-CPPs. As expected, the results confirmed a significant enrichment of positively charged residues like arginine (R) and lysine (K) in CPPs, consistent with previous studies.[20,32] Additionally, we observed that aromatic residues (tryptophan (W), tyrosine (Y)) and histidine (H) also showed noticeable differences between CPPs and non-CPPs (Figures 2 and S2). Conversely, non-CPPs exhibited higher percentages of acidic residues (aspartic acid (D) and glutamic acid (E)) and other nonpolar residues.

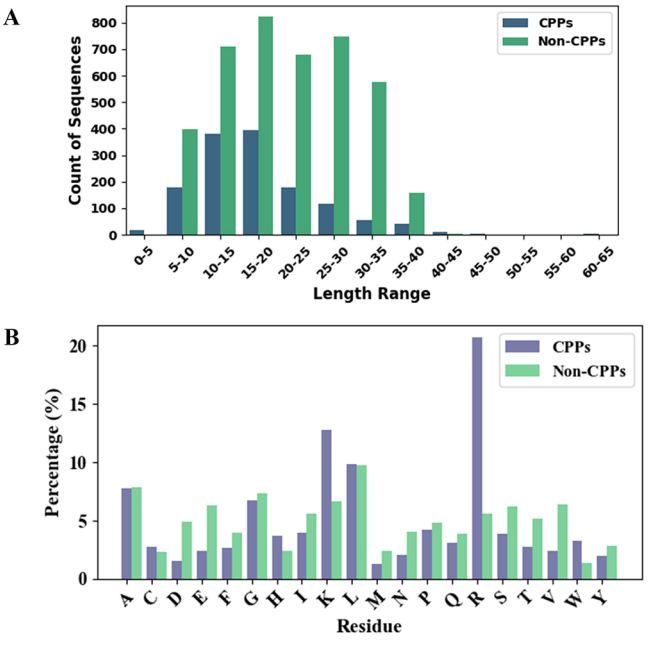


Figure 2. (A) Distribution of peptide sequence lengths for CPPs (cell-penetrating peptides) and non-CPPs. The plot illustrates the count of sequences across different length ranges. (B) Distribution of amino acid residues in CPP and non-CPP sequences represented as percentages. The bars show the relative abundance of each residue type among CPPs and non-CPPs, calculated from a total of 23,505 resides of 1380 CPP sequences and 84,918 residues of 4099 non-CPP sequences.

To gain deeper insights into the key features distinguishing CPPs from non-CPPs, we performed sequence logo analysis (Figure 3). This analysis generates a graphical representation of the amino acid composition at each position in a set of aligned sequences, highlighting conserved and variable residues. The figures are arranged in columns (top to bottom) and ranked based on their statistical significance determined by the MEME suite. The sequence logo analysis confirmed the significant prevalence of arginine (R) and lysine (K) residues, complementing the findings above and aligning with previous experimental results.[3,4] Additionally, the consistent presence of aromatic residues (tryptophan (W), tyrosine (Y)), as well as leucine (L), methionine (M), and phenylalanine (F) at various positions suggests their importance in differentiating CPPs from non-CPPs. These residues might contribute to stabilizing the peptide structure and interacting with cellular membranes.[57,58] In contrast, the analysis of non-CPPs revealed a more diverse and less consistent distribution of amino acids (Figure S3). While hydrophobic and aromatic residues were present, their patterns and positional preferences differed from those observed in CPPs (Figures 3 and S3). Overall, this comprehensive analysis suggests that CPPs share common features, including enrichment of positively charged residues (arginine and lysine), hydrophobic residues (leucine and phenylalanine), and aromatic residues (tryptophan and tyrosine). The conserved motifs and positional preferences of these residues highlight specific regions and residues critical for differentiating CPPs.[23,57−60]

3.2. Peptide Embedding Analysis. While the analysis of peptide sequences provided valuable insights into the compositional and length distribution patterns differentiating CPPs from non-CPPs, it primarily focused on the presence of specific amino acids. To delve deeper and capture the potentially intricate relationships and hidden patterns within

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

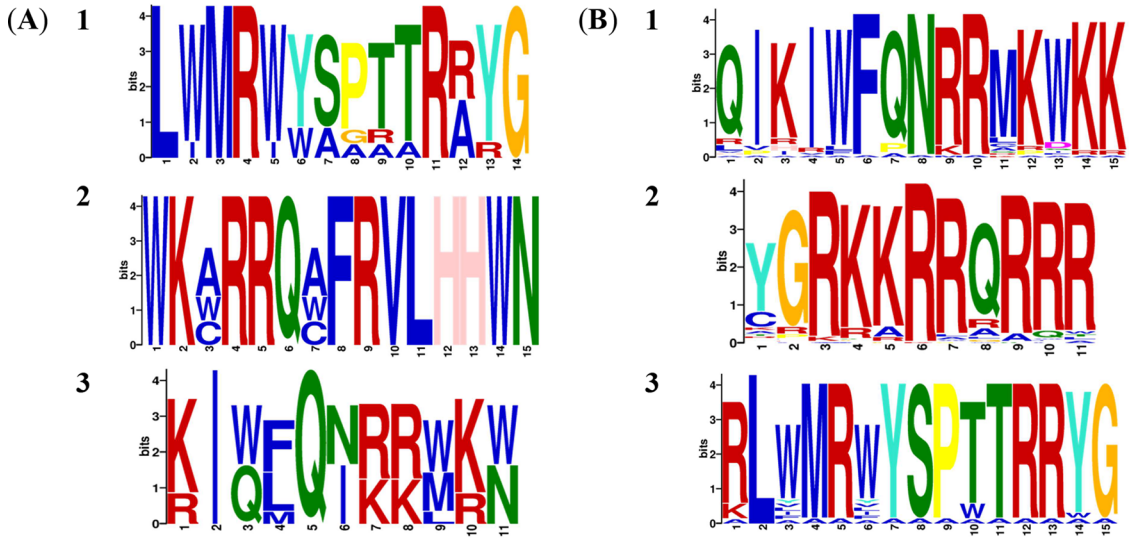


Figure 3. Sequence motif analysis of CPPs from the (A) kelm data set and (B) test and training data set, highlighting the most statistically significant motifs identified by the MEME suite. The logo plots are arranged in columns (1−3) and ranked based on their statistical significance (*E*-value) as determined by the MEME suite.

the sequences, we embedded the peptide sequences using various state-of-the-art PLMs. These models serve as the foundation for understanding the properties of the peptides and their subsequent classification tasks. The PLMs used in this study include BEPLER (121 features), CPCProt (512 features), SeqVec (1024 features), ESM2 (320, 480, 640, and 1280 features), ESM1b (1280 features), ESM1v (1280 features), ProtT5-XL UniRef50 (1024 features), ProtT5-XL BFD (1024 features), and ProtBERT (1024 features). These PLMs were trained with extensive protein sequence data, enabling them to learn rich representations of the sequences. This training process allows the models to capture contextual information and long-range dependencies within the peptide sequences, making them highly effective for generating informative embeddings. To visualize the distribution of these high-dimensional embeddings, we used UMAP, a dimensionality reduction technique specifically designed for data visualization. It prioritizes the preservation of local similarities between data points over global distances. This capability, along with its ability to handle outliers and nonlinear relationships, makes UMAP particularly effective for visualizing biological data compared to traditional methods like principal component analysis.[55] The UMAP distribution of positive and negative samples of training, test, and external data set was plotted in two-dimensional feature space created from all embeddings, as shown in Figures 2 and S1. The UMAP analysis revealed well-separated clusters for CPPs (positive) and non-CPPs (negative), suggesting that the embeddings effectively capture the inherent properties that differentiate these peptide sequences. This distinct separation in the two-dimensional space signifies a strong representation of peptide information, which is crucial for downstream model development and achieving optimal performance.[61,62] This aligns with the observed performance of pLM4CPPs on the benchmark data sets (Table 2). These findings pave the way for further analysis of the peptide embeddings and their potential for the CPP classification model development.

3.3. Performance of PLM-Based Models on the Test Data Set. We evaluated various embedding models on the test data set to gauge the generalization capabilities of each model. All models exhibited high ACC values exceeding 0.90, but BEPLER, CPCProt, and ProtBERT showed lower BACC values of 0.869, 0.853, and 0.871, respectively. These models also had the lowest MCC values, indicating less effective performance compared to other embeddings. SeqVec, on the other hand, showed the highest ACC (0.932) and BACC (0.901), suggesting its robust performance. Among the ESM2 variants, ESM2−480 performed superbly well, achieving an ACC of 0.931 and a BACC of 0.907, along with high sensitivity (0.860) and specificity (0.955), highlighting its balanced and comprehensive classification capabilities. ProtT5-XL UniRef50 demonstrated the highest specificity (0.977), indicating a strong ability to correctly identify negative cases. Despite the high ACC (0.927) of ProtBERT, its slightly lower Sn (0.799) and MCC (0.751) values suggest some limitations in its performance. The ESM2 models, particularly ESM2−480 and ESM2−1280, exhibited high MCC values (0.816 and 0.808, respectively), emphasizing their ability to manage the complexities of peptide sequences. The consistent AUC values across most models, with SeqVec achieving the highest (0.901), further underscore their capability in distinguishing between positive and negative cases. This detailed analysis highlights the variability in model performance, emphasizing the importance of selecting appropriate embedding models tailored to the specific characteristics of the peptide sequences to achieve optimal classification results. Overall, the performance of these all-embedding models suggests that models like SeqVec and ESM2 variants are particularly strong candidates for achieving high classification performance, while others like BEPLER and CPCProt might be less effective. These findings provide valuable insights into the strengths and characteristics of each embedding model, aiding in the selection of the most suitable model for classification and prediction of the biological activity of peptides.

3.4. Model Performance on the External kelm Data Set. Several publicly available ML-based Cell-Penetrating Peptide prediction models have been comprehensively reviewed covering the biological significance of CPPs and existing ML methods for CPP prediction.[34] To evaluate these prediction models, Su et al. conducted an empirical

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Comparison Performance Metrics Such as Accuracy (ACC), Balanced Accuracy (BACC), Sensitivity (Sn), Specificity (Sp), Matthews Correlation Coefficient (MCC), and Area under the ROC Curve (AUC) of Different Embedding Models on Test Dataset Using a CNN Classifier[a]

embeddings models	embedding dimension	ACC	BACC	Sn	Sp	MCC	AUC
BEPLER	121	0.908	0.869	0.791	0.947	0.752	0.869
CPCProt	512	0.902	0.853	0.752	0.954	0.735	0.853
SeqVec	1024	0.932	0.901	0.838	0.965	0.819	0.901
ESM2	320	0.923	0.892	0.831	0.955	0.795	0.893
ESM2	480	0.931	0.907	0.860	0.955	0.816	0.907
ESM2	640	0.923	0.880	0.791	0.968	0.792	0.880
ESM2	1280	0.929	0.893	0.820	0.966	0.808	0.892
ESM1b	1280	0.920	0.883	0.809	0.957	0.784	0.883
ESM1v	1280	0.923	0.889	0.820	0.958	0.794	0.889
ProtT5-XL UniRef50	1024	0.925	0.875	0.773	0.977	0.797	0.875
ProtT5-XL BFD	1024	0.921	0.891	0.831	0.951	0.789	0.891
ProtBERT	1024	0.927	0.871	0.799	0.944	0.751	0.871

[a]ACC and MCC were used to select the best-performing models.

comparison of 12 models from six publicly available CPP prediction tools on benchmark validation sets containing CPPs and non-CPPs from the kelm data set.[34] The kelm data set consists of 96 experimentally validated CPPs and 96 non-CPPs.[53] For the empirical comparison, sequences that did not meet the length requirements of certain predictors were removed from the data set as some servers impose strict length limitations for input sequences. For example, CellPPD limits sequences to 1−50 residues, SkipCPP-Pred to no less than 10 residues, and KELM-CPP-pred to 5−30 residues.[31,53,63] Additionally, they excluded the sequences with identity of >30% against the sequences in the training data sets during the empirical comparison.[34] This empirical comparison demonstrated that MLCPP performed well with these independent validation sets. Building upon it, Manavalan and Patra further improved MLCPP by utilizing a larger training data set, various sequence-derived features, and conventional ML classifiers, resulting in the next generation MLCPP 2.0.[25] To assess the performance of our models and benchmark them against existing methods, we first evaluated MLCPP 2.0 on the complete kelm data set, including all sequences without any exclusions based on length (Tables 1 and S1). This approach ensured a more comprehensive evaluation by avoiding potential biases introduced through sequence length restrictions. Additionally, unlike the empirical analysis by Su et al., we did not exclude sequences based on sequence homology with training data sets. This is because the peptide embeddings effectively capture the inherent properties that differentiate these peptide sequences, even with potential sequence similarities, as shown in Figure 4. By including all available sequences, we aimed to explore the robustness and generalizability of the models. The comparative results are presented in Table 2.

A cursory view at the performance metrics of our models suggests that most of the models outperform MLCPP 2.0, except for CPCProt and ProtBERT. Specifically, our ProtT5-XL BFD model showed exceptional performance on the kelm data set, achieving the highest accuracy (ACC = 0.901) among our models. It also demonstrated balanced Sn (0.885) and Sp (0.917), along with a high MCC (0.802), suggesting an excellent overall performance. ESM2−1280 model also achieved superior performance, particularly in terms of specificity (Sp = 0.978), accuracy (ACC = 0.896), and MCC (0.796). The high specificity indicates that our model effectively identifies non-CPP sequences, which is crucial for reducing false positives. The ACC and MCC values emphasize the balanced and reliable predictive power of these models compared with other state-of-the-art models. SeqVec, on the other hand, maintained an exceptional Sp (0.938), making it a valuable choice for tasks requiring a strong true negative prediction rate (identifying non-CPPs). Among the ESM2 variants, models with feature dimensions of 320, 480, and 640 displayed consistent performance with accuracy values around 0.880, high sensitivity (ranging from 0.802 to 0.900), and MCC values around 0.735, highlighting their robustness. The ESM-1b and ESM-1v models achieved accuracy values of 0.865 and 0.859, respectively, with MCC scores of 0.735 and 0.724. Both models maintained good Sn and Sp, with ESM-1b at 0.802 and 0.927, and ESM-1v at 0.802 and 0.917. ProtT5-XL UniRef50 achieved an accuracy of 0.875 with high Sp (0.948) and an MCC of 0.758, suggesting its ability to correctly identify negative cases. BEPLER achieved an ACC of 0.865 and an MCC of 0.737, with Sn and Sp values of 0.792 and 0.938, respectively. While BEPLER demonstrated a good balance, its slightly lower sensitivity suggests potential limitations in identifying true positive cases compared to other models. CPCProt exhibited a lower overall performance with an ACC of 0.833, an MCC of 0.668, and Sn and Sp values of 0.802 and 0.865, respectively. This model might require further development to handle diverse peptide sequences more effectively. Similarly, ProtBERT had an ACC of 0.833 and an MCC of 0.672, with sensitivity and specificity values of 0.771 and 0.896, respectively. Its lower Sn concentration suggests limitations in identifying true positives. In comparison, the MLCPP 2.0 model, which employs an ensemble learning approach that includes conventional ML classifiers with various sequence-based feature encoding algorithms, achieved an ACC of 0.854, a sensitivity of 0.740, and a Sp of 0.969, with an MCC of 0.728. While MLCPP 2.0 showed high Sp, its Sn was lower compared with our best-performing models. This highlights the advantage of our models in achieving a more balanced performance. Overall, our embedding models, particularly ESM2−1280 and ProtT5-XL BFD, demonstrated superior performance on the kelm data set, with high ACC, Sn, Sp, and MCC values.

This comprehensive evaluation of our models on test and external independent data set confirms the robustness and generalizability of models such as SeqVec, ESM2 variants, and

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

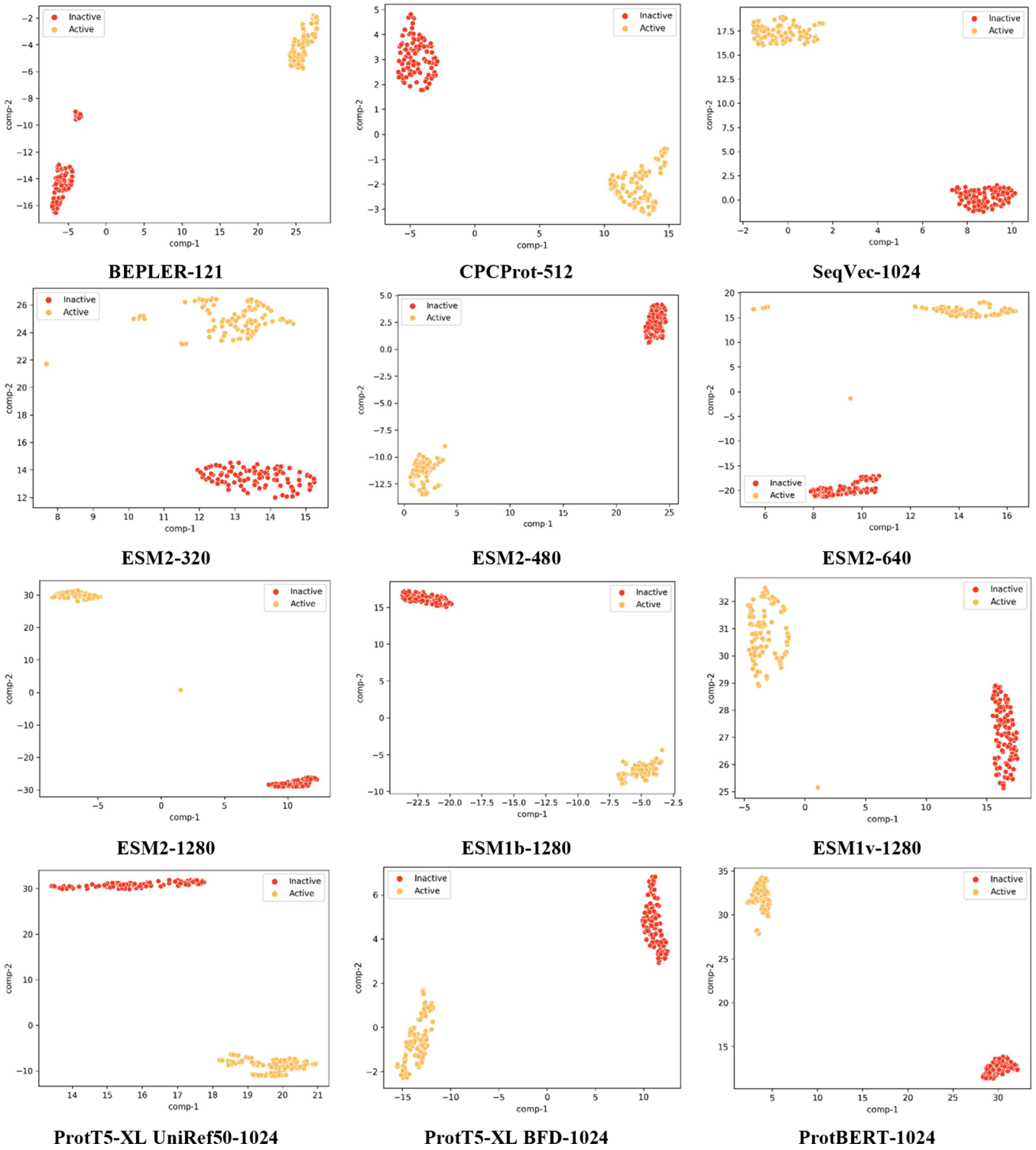


Figure 4. Uniform manifold approximation and projection (UMAP) visualization of active and inactive samples from the kelm data set embedded by various pretrained protein language models.

ProtT5-XL BFD across different data sets. Additionally, we conducted a 10-fold cross-validation to further assess the stability and reliability of these top performing models. The results, summarized in Table 3, show consistent performance across all folds, with high ACC, BACC, and MCC for the top-performing embeddings. These results indicate that our model generalizes well across different subsets of the data. The cross-validation analysis confirms that the SeqVec and ESM2 variants, as well as ProtT5-XL BFD embeddings, in combination with our CNN-based architecture, offer the most reliable performance for CPPs classification. Furthermore, we evaluated our models using five popular traditional classifiers: Logistic Regression, Random Forest, Support Vector Machine, k-Nearest Neighbors, and Multilayer Perceptron, using the same PLMs embeddings. It is worth noting that traditional classifiers have been reported to perform

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Performance Metrics Such as Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), Matthews Correlation Coefficient (MCC), True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) of Our Models and Recently Reported MLCPP 2.0 and Other Available Models on the kelm Dataset[a]

CNN classifier	features	TP	FP	TN	FN	Sn	Sp	ACC	MCC
our models									
BEPLER	121	76	6	90	20	0.792	0.938	0.865	0.737
CPCProt	512	77	13	83	19	0.802	0.865	0.833	0.668
SeqVec	1024	77	6	90	19	0.802	0.938	0.870	0.746
ESM-2	320	86	13	83	10	0.900	0.865	0.880	0.761
ESM-2	480	77	7	89	19	0.802	0.927	0.865	0.735
ESM-2	640	83	10	86	13	0.865	0.896	0.880	0.761
ESM-2	1280	81	5	91	15	0.844	0.978	0.896	0.796
ESM-1b	1280	77	7	89	19	0.802	0.927	0.865	0.735
ESM-1v	1280	77	8	88	19	0.802	0.917	0.859	0.724
ProtT5-XL UniRef50	1024	77	5	91	19	0.802	0.948	0.875	0.758
ProtT5-XL BFD	1024	85	8	88	11	0.885	0.917	0.901	0.802
ProtBERT	1024	74	10	86	22	0.771	0.896	0.833	0.672
models reported in the literature on the kelm data set									
MLCPP 2.0		71	3	93	25	0.740	0.969	0.854	0.728
MLCPP		53	5	43	18	0.747	0.896	0.808	0.630
CPPred-RF		59	12	36	12	0.831	0.750	0.798	0.580
KELM-AAC		49	5	43	22	0.690	0.896	0.773	0.580
KELM-hybrid-AAC		49	5	43	22	0.690	0.896	0.773	0.580
CPPred-FL		56	10	38	15	0.789	0.792	0.790	0.570
CellPPD		45	3	45	26	0.634	0.938	0.756	0.570
CellPPD-motif		45	3	45	26	0.634	0.938	0.756	0.570
KELM-PseAAC		59	13	35	12	0.831	0.729	0.790	0.560
KELM-DAC		40	1	47	31	0.563	0.979	0.731	0.560
SkipCPP-Pred		58	13	35	13	0.817	0.729	0.782	0.550
KELM-hybrid-PseAAC		59	14	34	12	0.831	0.708	0.782	0.540
KELM-hybrid-DAC		49	8	40	22	0.690	0.833	0.748	0.510

[a]ACC and MCC were used to select the best-performing models.

Table 3. Top Model Performance Metrics Such as Averaged Accuracy (ACC), Balanced Accuracy (BACC), Sensitivity (Sn), Specificity (Sp), Matthews Correlation Coefficient (MCC), and Area under the ROC Curve (AUC) Were Obtained from 10-Fold Cross-Validation Using CNN Classifiers for the Top-Performing Models

embeddings models	embedding dimension	ACC	BACC	Sn	Sp	MCC	AUC
SeqVec	1024	0.931 ± 0.009	0.896 ± 0.016	0.825 ± 0.034	0.968 ± 0.008	0.816 ± 0.024	0.949 ± 0.018
ESM2	320	0.931 ± 0.008	0.898 ± 0.017	0.831 ± 0.038	0.965 ± 0.010	0.816 ± 0.024	0.951 ± 0.010
ESM2	480	0.927 ± 0.013	0.885 ± 0.029	0.800 ± 0.623	0.970 ± 0.008	0.803 ± 0.038	0.944 ± 0.016
ESM2	640	0.925 ± 0.005	0.885 ± 0.009	0.803 ± 0.018	0.967 ± 0.009	0.799 ± 0.017	0.947 ± 0.015
ESM2	1280	0.928 ± 0.009	0.889 ± 0.015	0.838 ± 0.034	0.959 ± 0.014	0.809 ± 0.028	0.950 ± 0.012
ProtT5-XL BFD	1024	0.943 ± 0.006	0.885 ± 0.145	0.807 ± 0.034	0.963 ± 0.008	0.794 ± 0.018	0.943 ± 0.013

comparably to deep learning methods in some cases, such as the classification of antimicrobial peptides.[64] The comparison results, summarized in Table S2, indicate that while traditional classifiers perform respectably, the CNN model consistently achieves higher performance across most metrics, particularly in terms of MCC and ACC. For instance, with SeqVec embeddings, the CNN model achieved an MCC of 0.819 and an ACC of 0.932, outperforming the best traditional model (Logistic Regression), which had an MCC of 0.777 and an ACC of 0.912. Similarly, for other embeddings like ESM2 and ProtT5-XL BFD, the CNN model maintained superior performance compared with traditional classifiers. Additionally, Our CNN models demonstrated superior performance compared to traditional classifiers (for instance, XGBoost) on the external kelm data set. Among the CNN models, ProtT5-XL BFD (1024 features) and ESM2 variants (except for 320 features) showed the best ACC, sensitivity, and MCC. In comparison, XGBoost models exhibited slightly lower MCC values, with the best performance achieved by ESM2 (1280 features) (MCC: 0.744) and ProtT5-XL BFD (1024 features) (MCC: 0.702). However, the XGBoost models showed higher specificity across all embeddings but were less effective in sensitivity compared to CNN models, suggesting CNN-based models consistently outperformed XGBoost, as shown in Table 4.

Furthermore, we evaluated the performance of the k-NN classifier using the 320-dimensional ESM-2 embeddings and observed competitive results. Specifically, k-NN achieved a slightly higher AUC (0.951) and MCC (0.800) than the CNN model on the test data set, along with a better specificity (0.967). However, on the external kelm data set, the CNN model outperformed k-NN in balanced accuracy, sensitivity, and MCC, indicating better generalization across unseen data, as shown in Table S3. While CNN models demonstrated

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 4. Comparison of Performance Metrics, Including Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), Matthews Correlation Coefficient (MCC), True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), between our CNN Models and Traditional Classifier Models Such as XGBoost on the KELM Dataset

CNN classifier	features	TP	FP	TN	FN	Sn	Sp	ACC	MCC
our models									
SeqVec	1024	76	6	90	20	0.792	0.938	0.865	0.737
ESM2	320	77	13	83	19	0.802	0.865	0.833	0.668
ESM2	480	77	6	90	19	0.802	0.938	0.870	0.746
ESM2	640	86	13	83	10	0.900	0.865	0.880	0.761
ESM2	1280	77	7	89	19	0.802	0.927	0.865	0.735
ProtT5-XL BFD	1024	83	10	86	13	0.865	0.896	0.880	0.761
traditional classifiers									
XGBoost									
SeqVec	1024	66	3	93	30	0.688	0.969	0.828	0.684
ESM2	320	68	3	93	28	0.708	0.969	0.839	0.701
ESM2	480	65	2	94	31	0.677	0.979	0.828	0.688
ESM2	640	71	3	93	25	0.740	0.969	0.854	0.728
ESM2	1280	70	1	95	26	0.729	0.900	0.859	0.744
ProtT5-XL BFD	1024	65	1	95	31	0.677	0.990	0.833	0.702

superior performance across most embeddings and metrics, the results highlight that a simpler model, such as k-NN, can be competitive in such classification tasks. Future work will explore systematic comparisons of simpler models (e.g., k-NN, decision trees) with deep learning approaches across diverse embeddings to assess their interpretability, performance trade-offs, and utility for different data sets. This exploration could guide the development of models that better balance the predictive power and explainability. However, it is crucial to highlight that regardless of the classifier, the use of PLMs significantly enhances feature representations, demonstrating their importance in improving model performance. While traditional classifiers and k-NN have their merits, the CNN model clearly demonstrates a performance advantage for our data set in terms of predictive power. These comprehensive analyses and comparison emphasize the importance of selecting models that perform well not only on test data sets but also on external data sets, ensuring their applicability and reliability for predicting CPPs. Additionally, it also suggests that using feature representations based on PLMs may be a potential solution to the feature representation challenges commonly faced in developing ML-based models. While the current approach focuses on sequence-based embeddings without incorporating spatial information, future studies will aim to further refine the methodology to enhance its robustness and broader applicability.

4. CONCLUSIONS

The field of bioinformatics offers immense potential to significantly reduce the time and cost associated with exploring novel bioactive peptides. Accurate and rapid prediction models are crucial for this advancement. AMPs represent a diverse class of molecules with broad therapeutic potential. CPPs, a subset of AMPs, possess the unique ability to deliver cargo directly into cells, making them valuable tools for drug delivery and gene therapy applications. In this study, we introduce pLM4CCPs, a state-of-the-art deep learning model that leverages pLMs for peptide embedding and CNNs for classifying CPPs. To our knowledge, this work presents the most comprehensive evaluation of various pLMs for CPPs classification, including BEPLER, CPCProt, SeqVec, ESM variants, ProtT5 models, and ProtBERT. Our findings highlight the superior performance of ESM-1280 and ProtT5-XL BFD embeddings in representing CPPs, achieving high accuracy and reliability. The pLM4CCP model, employing CNNs for classification, demonstrates notable improvements over the existing state-of-the-art CPP prediction methods. Specifically, pLM4CCPs achieve significant enhancements in ACC (4.9−5.5%), MCC (9.3−10.2%), and Sn (14.1−19.6%). ESM-1280 achieved an ACC of 0.896, an MCC of 0.796, an Sn of 0.844, and an Sp of 0.978. Similarly, the ProtT5-XL BFD achieved an accuracy of 0.901, MCC of 0.802, Sn of 0.885, and Sp of 0.917. These results underscore the efficacy of these embeddings in capturing the essential features for accurate CPP classification. The approaches integrated into pLM4CCPs, which consolidate predictions from multiple models, further enhance the reliability of peptide classification. This methodology leverages the individual strengths of each model, providing a robust consensus that significantly strengthens the prediction accuracy and reliability. To benefit the research community, we have developed a user-friendly web server for bioactivity predictions, available at https://ry2acnp6ep.us-east-1.awsapprunner.com. The related source code, data sets, and adaptable templates, including resources for embedding generation, model training, evaluation, and protocols for predicting peptide activity using pLM4CPP models, are freely available on GitHub at https://github.com/drkumarnandan/pLM4CPPs. This resource is intended to support further research and development in peptide functionality and classification, enabling researchers to explore and validate peptide-based applications more effectively. Additionally, our approach and architecture can be transferred to other bioactive peptide predictions beyond CPPs, demonstrating its versatility and broad applicability in the field.

■ ASSOCIATED CONTENT

Data Availability Statement

All source code and data sets used in this publication are freely available for academic use under an MIT license at https://github.com/drkumarnandan/pLM4CPPs. The training and testing data sets are in the "data set" directory. Additionally, the independent evaluation data set is also stored in the "data set" directory.

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c01338.

Figures S1 and S2 show the amino acid distribution and lengths of CPP and non-CPP sequences. Figure S3 shows sequence motif analysis of non-CPP sequences. Figure S4 shows UMAP visualization of positive and negative samples of used data set. Table S1 shows evaluation of external data set. Table S2 shows performance metrics comparison of the best models on the test data set with CNN and six popular traditional classifiers. Table S3 shows performance metrics comparison of ESM-320 embeddings on test and external data sets using CNN and k-NN classifier (PDF)

AUTHOR INFORMATION

Corresponding Author

Yonghui Li − *Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0003-4320-0806; Phone: 785-532-4061; Email: yonghui@ksu.edu; Fax: 785-532-4061

Authors

Nandan Kumar − *Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0001-6915-708X

Zhenjiao Du − *Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0002-8492-4328

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c01338

Author Contributions

N.K.: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Visualization, Writing - Original Draft, Writing - Review and Editing. Z.D.: Methodology, Formal analysis, Investigation, Writing - Review and Editing. Y.L.: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing - Review and Editing, Supervision, Project administration, and Funding acquisition

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This is contribution no. 25-020-J from the Kansas Agricultural Experimental Station.

REFERENCES

(1) Gao, X.; Kim, K.-S.; Liu, D. Nonviral Gene Delivery: What We Know and What Is Next. *AAPS J.* **2007**, *9* (1), E92−104.

(2) Walther, W.; Stein, U. Viral Vectors for Gene Transfer: A Review of Their Use in the Treatment of Human Diseases. *Drugs* **2000**, *60* (2), 249−271.

(3) Glover, D. J.; Lipps, H. J.; Jans, D. A. Towards Safe, Non-Viral Therapeutic Gene Expression in Humans. *Nat. Rev. Genet* **2005**, *6* (4), 299−310.

(4) Järver, P.; Langel, Ü. The Use of Cell-Penetrating Peptides as a Tool for Gene Regulation. *Drug Discovery Today* **2004**, *9* (9), 395−402.

(5) Kumar, N.; Sastry, G. N. Study of Lipid Heterogeneity on Bilayer Membranes Using Molecular Dynamics Simulations. *Journal of Molecular Graphics and Modelling* **2021**, *108*, No. 108000.

(6) Chen, N.; Jiang, C. Antimicrobial Peptides: Structure, Mechanism, and Modification. *Eur. J. Med. Chem.* **2023**, *255*, No. 115377.

(7) Galzitskaya, O. V. Creation of New Antimicrobial Peptides. *International Journal of Molecular Sciences* **2023**, *24* (11), 9451.

(8) Mhlongo, J. T.; Waddad, A. Y.; Albericio, F.; de la Torre, B. G. Antimicrobial Peptide Synergies for Fighting Infectious Diseases. *Advanced Science* **2023**, *10* (26), No. 2300472.

(9) Madani, F.; Lindberg, S.; Langel, Ü.; Futaki, S.; Gräslund, A. Mechanisms of Cellular Uptake of Cell-Penetrating Peptides. *J. Biophys.* **2011**, *2011*, No. e414729.

(10) Milletti, F. Cell-Penetrating Peptides: Classes, Origin, and Current Landscape. *Drug Discov Today* **2012**, *17* (15−16), 850−860.

(11) Qian, Z.; LaRochelle, J. R.; Jiang, B.; Lian, W.; Hard, R. L.; Selner, N. G.; Luechapanichkul, R.; Barrios, A. M.; Pei, D. Early Endosomal Escape of a Cyclic Cell-Penetrating Peptide Allows Effective Cytosolic Cargo Delivery. *Biochemistry* **2014**, *53* (24), 4034−4046.

(12) Margus, H.; Padari, K.; Pooga, M. Cell-Penetrating Peptides as Versatile Vehicles for Oligonucleotide Delivery. *Mol. Ther* **2012**, *20* (3), 525−533.

(13) Lehto, T.; Kurrikoff, K.; Langel, Ü. Cell-Penetrating Peptides for the Delivery of Nucleic Acids. *Expert Opin Drug Deliv* **2012**, *9* (7), 823−836.

(14) Presente, A.; Dowdy, S. F. PTD/CPP Peptide-Mediated Delivery of siRNAs. *Curr. Pharm. Des.* **2013**, *19* (16), 2943−2947.

(15) Bendifallah, N.; Rasmussen, F. W.; Zachar, V.; Ebbesen, P.; Nielsen, P. E.; Koppelhus, U. Evaluation of Cell-Penetrating Peptides (CPPs) as Vehicles for Intracellular Delivery of Antisense Peptide Nucleic Acid (PNA). *Bioconjug Chem.* **2006**, *17* (3), 750−758.

(16) Boisguerin, P.; Giorgi, J.-M.; Barrère-Lemaire, S. CPP-Conjugated Anti-Apoptotic Peptides as Therapeutic Tools of Ischemia-Reperfusion Injuries. *Curr. Pharm. Des* **2013**, *19* (16), 2970−2978.

(17) Hansen, A.; Schäfer, I.; Knappe, D.; Seibel, P.; Hoffmann, R. Intracellular Toxicity of Proline-Rich Antimicrobial Peptides Shuttled into Mammalian Cells by the Cell-Penetrating Peptide Penetratin. *Antimicrob. Agents Chemother.* **2012**, *56* (10), 5194−5201.

(18) Nasrollahi, S. A.; Fouladdel, S.; Taghibiglou, C.; Azizi, E.; Farboud, E. S. A Peptide Carrier for the Delivery of Elastin into Fibroblast Cells. *Int. J. Dermatol* **2012**, *51* (8), 923−929.

(19) Xia, H.; Gao, X.; Gu, G.; Liu, Z.; Hu, Q.; Tu, Y.; Song, Q.; Yao, L.; Pang, Z.; Jiang, X.; Chen, J.; Chen, H. Penetratin-Functionalized PEG-PLA Nanoparticles for Brain Drug Delivery. *Int. J. Pharm.* **2012**, *436* (1−2), 840−850.

(20) Heitz, F.; Morris, M. C.; Divita, G. Twenty Years of Cell-Penetrating Peptides: From Molecular Mechanisms to Therapeutics. *Br. J. Pharmacol.* **2009**, *157* (2), 195−206.

(21) Altelaar, A. F. M.; Munoz, J.; Heck, A. J. R. Next-Generation Proteomics: Towards an Integrative View of Proteome Dynamics. *Nat. Rev. Genet* **2013**, *14* (1), 35−48.

(22) Rhodes, C. A.; Pei, D. Bicyclic Peptides as Next-Generation Therapeutics. *Chemistry* **2017**, *23* (52), 12690−12703.

(23) Dougherty, P. G.; Sahni, A.; Pei, D. Understanding Cell Penetration of Cyclic Peptides. *Chem. Rev.* **2019**, *119* (17), 10241−10287.

(24) Ragin, A. D.; Morgan, R. A.; Chmielewski, J. Cellular Import Mediated by Nuclear Localization Signal Peptide Sequences. *Chem. Biol.* **2002**, *9* (8), 943−948.

(25) Manavalan, B.; Patra, M. C. MLCPP 2.0: An Updated Cell-Penetrating Peptides and Their Uptake Efficiency Predictor. *J. Mol. Biol.* **2022**, *434* (11), No. 167604.

(26) Qiu, X.; Li, H.; Ver Steeg, G.; Godzik, A. Advances in AI for Protein Structure Prediction: Implications for Cancer Drug Discovery and Development. *Biomolecules* **2024**, *14* (3), 339.

(27) Diener, C.; Martínez, G. G. R.; Blas, D. M.; González, D. A. C.; Corzo, G.; Castro-Obregon, S.; Rio, G. D. Effective Design of Multifunctional Peptides by Combining Compatible Functions. *PLoS Comput. Biol.* **2016**, *12* (4), No. e1004786.

(28) Holton, T. A.; Pollastri, G.; Shields, D. C.; Mooney, C. CPPpred: Prediction of Cell Penetrating Peptides. *Bioinformatics* **2013**, *29* (23), 3094−3096.
(29) Gautam, A.; Chaudhary, K.; Kumar, R.; Raghava, G. P. S. Computer-Aided Virtual Screening and Designing of Cell-Penetrating Peptides. *Methods Mol. Biol.* **2015**, *1324*, 59−69.
(30) Tang, H.; Su, Z.-D.; Wei, H.-H.; Chen, W.; Lin, H. Prediction of Cell-Penetrating Peptides with Feature Selection Techniques. *Biochem. Biophys. Res. Commun.* **2016**, *477* (1), 150−154.
(31) Wei, L.; Tang, J.; Zou, Q. SkipCPP-Pred: An Improved and Promising Sequence-Based Predictor for Predicting Cell-Penetrating Peptides. *BMC Genomics* **2017**, *18* (Suppl 7), 742.
(32) Wei, L.; Xing, P.; Su, R.; Shi, G.; Ma, Z. S.; Zou, Q. CPPred-RF: A Sequence-Based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. *J. Proteome Res.* **2017**, *16* (5), 2044−2053.
(33) de Oliveira, E. C. L.; Santana, K.; Josino, L.; Lima E Lima, A. H.; de Souza de Sales Júnior, C. Predicting Cell-Penetrating Peptides Using Machine Learning Algorithms and Navigating in Their Chemical Space. *Sci. Rep* **2021**, *11* (1), 7628.
(34) Su, R.; Hu, J.; Zou, Q.; Manavalan, B.; Wei, L. Empirical Comparison and Analysis of Web-Based Cell-Penetrating Peptide Prediction Tools. *Briefings in Bioinformatics* **2020**, *21* (2), 408−420.
(35) García-Jacas, C. R.; García-González, L. A.; Martinez-Rios, F.; Tapia-Contreras, I. P.; Brizuela, C. A. Handcrafted versus Non-Handcrafted (Self-Supervised) Features for the Classification of Antimicrobial Peptides: Complementary or Redundant? *Briefings in Bioinformatics* **2022**, *23* (6), bbac428.
(36) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16* (12), 1315−1322.
(37) Du, Z.; Ding, X.; Hsu, W.; Munir, A.; Xu, Y.; Li, Y. pLM4ACE: A Protein Language Model Based Predictor for Antihypertensive Peptide Screening. *Food Chem.* **2024**, *431*, No. 137162.
(38) Du, Z.; Ding, X.; Xu, Y.; Li, Y. UniDL4BioPep: A Universal Deep Learning Architecture for Binary Classification in Peptide Bioactivity. *Briefings in Bioinformatics* **2023**, *24* (3), bbad135.
(39) Du, Z.; Xu, Y.; Liu, C.; Li, Y. pLM4Alg: Protein Language Model-Based Predictors for Allergenic Proteins and Peptides. *J. Agric. Food Chem.* **2024**, *72* (1), 752−760.
(40) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *44* (10), 7112−7127.
(41) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123−1130.
(42) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (15), No. e2016239118.
(43) Wang, L.; Niu, D.; Zhao, X.; Wang, X.; Hao, M.; Che, H. A Comparative Analysis of Novel Deep Learning and Ensemble Learning Models to Predict the Allergenicity of Food Proteins. *Foods* **2021**, *10* (4), 809.
(44) Martínez-Mauricio, K. L.; García-Jacas, C. R.; Cordoves-Delgado, G. Examining Evolutionary Scale Modeling-Derived Different-Dimensional Embeddings in the Antimicrobial Peptide Classification through a KNIME Workflow. *Protein Sci.* **2024**, *33* (4), No. e4928.
(45) Cordoves-Delgado, G.; García-Jacas, C. R. Predicting Antimicrobial Peptides Using ESMFold-Predicted Structures and ESM-2-Based Amino Acid Features with Graph Deep Learning. *J. Chem. Inf. Model.* **2024**, *64* (10), 4310−4321.
(46) Bepler, T.; Berger, B. Learning Protein Sequence Embeddings Using Information from Structure. *arXiv* October 16, 2019. .
(47) Lu, A. X.; Zhang, H.; Ghassemi, M.; Moses, A. Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization. *bioRXiv* September 6, 2020. .
(48) Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; Rost, B. Modeling Aspects of the Language of Life through Transfer-Learning Protein Sequences. *BMC Bioinformatics* **2019**, *20* (1), 723.
(49) Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function. *bioRxiv* November 17, 2021. .
(50) Rao, R.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J. F.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. *bioRxiv* July 13, 2021. .
(51) Iovino, B. G.; Ye, Y. Protein Embedding Based Alignment. *BMC Bioinformatics* **2024**, *25* (1), 85.
(52) Agrawal, P.; Bhalla, S.; Usmani, S. S.; Singh, S.; Chaudhary, K.; Raghava, G. P. S.; Gautam, A. CPPsite 2.0: A Repository of Experimentally Validated Cell-Penetrating Peptides. *Nucleic Acids Res.* **2016**, *44* (D1), D1098−D1103.
(53) Pandey, P.; Patel, V.; George, N. V.; Mallajosyula, S. S. KELM-CPPpred: Kernel Extreme Learning Machine Based Prediction Model for Cell-Penetrating Peptides. *J. Proteome Res.* **2018**, *17* (9), 3214−3222.
(54) Dallago, C.; Schütze, K.; Heinzinger, M.; Olenyi, T.; Littmann, M.; Lu, A. X.; Yang, K. K.; Min, S.; Yoon, S.; Morton, J. T.; Rost, B. Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. *Current Protocols* **2021**, *1* (5), No. e113.
(55) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* September 17, 2020. .
(56) Bailey, T. L.; Johnson, J.; Grant, C. E.; Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **2015**, *43* (W1), W39−W49.
(57) Chan, D. I.; Prenner, E. J.; Vogel, H. J. Tryptophan- and Arginine-Rich Antimicrobial Peptides: Structures and Mechanisms of Action. *Biochimica et Biophysica Acta (BBA) -. Biomembranes* **2006**, *1758* (9), 1184−1202.
(58) Wang, C.; Dong, S.; Zhang, L.; Zhao, Y.; Huang, L.; Gong, X.; Wang, H.; Shang, D. Cell Surface Binding, Uptaking and Anticancer Activity of L-K6, a Lysine/Leucine-Rich Peptide, on Human Breast Cancer MCF-7 Cells. *Sci. Rep* **2017**, *7* (1), 8293.
(59) Sayers, E. J.; Cleal, K.; Eissa, N. G.; Watson, P.; Jones, A. T. Distal Phenylalanine Modification for Enhancing Cellular Delivery of Fluorophores, Proteins and Quantum Dots by Cell Penetrating Peptides. *J. Controlled Release* **2014**, *195*, 55−62.
(60) Schmidt, N.; Mishra, A.; Lai, G. H.; Wong, G. C. L. Arginine-Rich Cell-Penetrating Peptides. *FEBS Lett.* **2010**, *584* (9), 1806−1813.
(61) Manavalan, B.; Basith, S.; Shin, T. H.; Wei, L.; Lee, G. mAHTPred: A Sequence-Based Meta-Predictor for Improving the Prediction of Anti-Hypertensive Peptides Using Effective Feature Representation. *Bioinformatics* **2019**, *35* (16), 2757−2765.
(62) Wei, L.; Hu, J.; Li, F.; Song, J.; Su, R.; Zou, Q. Comparative Analysis and Prediction of Quorum-Sensing Peptides Using Feature Representation Learning and Machine Learning Algorithms. *Brief. Bioinform.* **2020**, *21* (1), 106−119.
(63) Gautam, A.; Chaudhary, K.; Kumar, R.; Sharma, A.; Kapoor, P.; Tyagi, A.; Raghava, G. P. S. Open source drug discovery consortium. In Silico Approaches for Designing Highly Effective Cell Penetrating Peptides. *Journal of Translational Medicine* **2013**, *11* (1), 74.
(64) García-Jacas, C. R.; Pinacho-Castellanos, S. A.; García-González, L. A.; Brizuela, C. A. Do Deep Learning Models Make a Difference in the Identification of Antimicrobial Peptides? *Briefings in Bioinformatics* **2022**, *23* (3), bbac094.

https://doi.org/10.1021/acs.jcim.4c01338
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX