PepBERT: Lightweight language models for peptide representation

Zhenjiao Du, Yonghui Li

Department of Grain Science and Industry, Kansas State University, Manhattan, KS 66506, USA

Abstract

Protein language models (pLMs) have been widely adopted for various protein and peptide-related downstream tasks and demonstrated promising performance. However, short peptides are significantly underrepresented in commonly used pLM training datasets. For example, only 2.8% of sequences in the UniProt Reference Cluster (UniRef) contain fewer than 50 residues, which potentially limit the effectiveness of pLMs for peptide-specific applications. Here, we present PepBERT, a lightweight and efficient peptide language model specifically designed for encoding peptide sequences. Two versions of the model, PepBERT-large (4.9 million parameters) and PepBERT-small (1.86 million parameters), were pretrained from scratch using four custom peptide datasets and evaluated on nine peptide-related downstream prediction tasks. Both PepBERT models achieved performance superior or comparable to the benchmark model, ESM-2 with 7.5 million parameters, on 8 out of 9 datasets. Overall, PepBERT provides a compact yet effective solution for generating high-quality peptide representations for downstream applications such as bioactive peptide screening and drug discovery. The datasets, source codes, pretrained models, and tutorials for usage of PepBERT are available at https://github.com/dzjxzyd/PepBERT-large.

Keywords: Scientific large language model, peptide representation, drug discovery, machine learning, bioactive peptide

1. Introduction

Bioactive peptides—protein fragments that exhibit beneficial biological activities such as antioxidant, anticancer, and antimicrobial functions—are of considerable interest in nutraceutical, therapeutic and industrial applications [1,2]. Traditional wet-lab methods for peptide discovery are labor-intensive, expensive, and time-consuming; thus, machine learning approaches have become popular for virtual screening at a large scale [3]. A key step in developing effective machine learning models for peptide discovery is obtaining robust numerical representations of peptides [1].

With the advent of transformer architectures, large language models (LLMs) have achieved remarkable success in text representation by leveraging self-supervised learning on massive datasets [4]. Biological sequences (e.g., DNA, RNA, proteins) share a similar sequential structure with natural language, making them amenable to such techniques [4]. For instance, ESM-2 [5] is a prominent protein language model (pLM) that has been applied to a variety of prediction tasks including structure prediction [5], solubility [6], function annotation [7], etc.

UniRef as a comprehensive and non-redundant sequence database, has been widely used for pLM pretraining [8]. However, the scarcity of short peptide sequences in training data—only 2.8% of sequences in UniRef are shorter than 50 residues—may limit their effectiveness for peptide-specific tasks. Moreover, employing a full-scale protein language model for peptides may be unnecessarily complex and computationally intensive given the shorter sequence lengths involved. To address these challenges, we developed PepBERT—a lightweight peptide language model tailored specifically for peptide representation. Previous work, PeptideBERT [9], has contributed to peptide language model development. However, it focused on fine-tuning the pretrained ProteinBERT [10] for individual downstream tasks, rather than training a peptide-specific language model from scratch or fine-tuning a general-purpose peptide language model to enable broader representation learning across the peptide discovery community. To date, there is no dedicated pretrained peptide language model specifically for embeddings of short amino acid sequences.

In this study, we adopted bidirectional encoder representations from transformers (BERT)-based architecture with a masked language modeling strategy and pretrained our models on four datasets comprising between 1.97 and 19.2 million peptide sequences. The pretrained PepBERT models was released at Hugging Face

(https://huggingface.co/dzjxzyd) with detailed tutorals for usage in peptide embeddings. We then evaluated their performance on nine bioactive peptide datasets via transfer learning. Compared with the smallest variant of ESM-2 (7.5 million parameters), our models demonstrate equal or superior performance on 8 out of 9 datasets.

2. Methods

2.1 Datasets

Four datasets were compiled in September 2024. Three were extracted from the UniRef100, UniRef90, and UniRef50 databases, and one dataset was collected from the UniProt Archive (UniParc). All the sequences are unique in each dataset and shorter than 51 residues. The four datasets are summarized in Table 1 and are available for downloading on Hugging Face. The detailed length distributions of the datasets are illustrated in Figure 1.

2.2 Model architecture and pretraining setup

The overall mode architecture was based on Robustly Optimized BERT (RoBERTa), and the hyperparameters are referred to ESM-2 model with a few modifications [5,11]. Since all peptides are 50 residues or shorter, we directly employed sinusoidal positional embeddings. Two architectures were developed: PepBERT-small with 1.86 million parameters and 160 dimensional embeddings and PepBERT-large with 4.94 million parameters and 320 dimensional embeddings. Both models consist of 8 encoder layers, each featuring an 8-head multi-head self-attention mechanism. The feed-forward network in each layer uses an intermediate dimension of 640.

For optimization, we used the AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\varepsilon = 10^{-8}$, and a weight decay of 0.01. The learning rate was warmed up over the first 2000 steps to a peak of 4×10^{-4} and then decayed linearly to one-tenth of the peak value over 90% of the training duration. The maximum input length was fixed at 50 amino acids. During tokenization, the special tokens [SOS], [EOS], [PAD], [UNK], and [MASK] were incorporated. For the masked language modeling task, 15% of amino acids in each peptide were randomly replaced by the [MASK] token, and cross-entropy loss was computed over the predictions.

A validation set comprising 0.1% of the sequences was held out, and model checkpoints were saved after each epoch. The checkpoint with the lowest validation loss over 20 epochs was selected as the final model for each architecture and dataset. The training scripts are available at https://github.com/dzjxzyd/PepBERT-small and

https://github.com/dzjxzyd/PepBERT-large. A total of eight pretrained models can be downloaded from Hugging Face at https://huggingface.co/dzjxzyd, along with a tutorial for generating embeddings.

2.3 Transfer learning on bioactive peptide datasets and performance comparison

Nine bioactive peptide datasets (Table S1) were selected from previous studies in our laboratory [12,13]], restricting to sequences shorter than 50 residues. The pretrained PepBERT models were used to generate peptide embeddings for transfer learning. For benchmarking, we compared with the smallest variant of ESM-2—a well-established protein language model with 7.5 million parameters and 320-dimensional output embeddings [5]. Each dataset was randomly shuffled and split into training and test sets (80:20 ratio), repeated ten times. To ensure a fair comparison of the embeddings and to minimize the influence of the classifier itself, logistic regression (max_iter=1000) was used for classification model development. Statistical comparisons between ESM-2 and PepBERT performances were conducted using a t-test ($p < 0.05$), with evaluation metrics including accuracy (ACC), balanced accuracy (BACC), sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and area under the curve (AUC).

3. Results and discussion

In Figure 2, we can clearly observe the constant improvements to loss at validation set with the increased dataset size for both models. Specifically, the PepBERT trained on UniParc with 19.2 million peptide sequences had the lowest loss in PepBERT-small and PepBERT-large, respectively. Similar observation was also reported in other large language model development studies [14,15]. Based on Chinchilla scaling laws: 20 tokens-per-parameter ratios as the optimal for foundation model development, PepBERT-large and PepBERT-small needed 98 million and 37.2 million tokens, respectively, which was fulfilled in most cases in our study (Table 1). We believed that the improvements mostly attributed to the increasing diversity of the dataset. UniRef database is the clustered sets of sequences from UniProt Knowledgebase (UniProtKB), and UniRef100, UniRef90 and UniRef50 are filtered at varying sequence identity thresholds (100%, 90%, 50%). In contrast, UniParc is a comprehensive archive of protein sequences that acts as a historical record [8]. Larger datasets, particularly the UniParc dataset (Figure 1), preserved more peptides with only a few residues (Figure 1). Additionally, PepBERT-larger, which has more learnable parameters, outperformed PepBERT-small. This suggests that the learning capacity of PepBERT-small may be insufficient to capture the patterns in such large datasets.

The checkpoints corresponding to the lowest validation loss were then used for downstream transfer learning tasks, alongside the benchmark ESM-2 model. As shown in Tables 2 and 3, the models pretrained on UniParc consistently achieved the best performance among the four datasets. Specifically, compared with ESM-2, PepBERT-large-UniParc outperformed it in 3 of 9 datasets ($p < 0.05$) and reached equivalent performance in 5 of 9 datasets with no significant differences in balanced accuracy (BACC). In contrast, PepBERT-small-UniParc demonstrated significantly better performance in 1 of 9 datasets ($p < 0.05$) and achieved comparable performance in 7 of 9 datasets. Detailed performance metrics are provided in Table S2.

It is worth noting that while PepBERT is designed to generate improved representations for short amino acid sequences (i.e., peptides), there are still underrepresented groups, particularly peptides with 2 to 5 residues. These ultrashort peptides are especially important in peptide based small molecule drug development [2]. However, training a LLM using masked language modeling or next-token prediction is inherently challenging on those extremely short sequences considering the mechanisms behind LLMs. There is less contextual information for those short peptides and hence the quality of representations generated by LLM for these peptides may be limited, and expectations should be moderated accordingly. One potential solution is to adopt alternative linear notations—such as the simplified molecular-input line-entry system (SMILES) or self-referencing embedded strings (SELFIES) [16]—to transform shorter amino acid sequences into longer, text-like formats. These alternative notations also facilitate the extension to modified or cyclic peptides. Wang et al. recently introduced PepDoRA, a peptide representation model based on SMILES notation [17]. PepDoRA fine-tunes ChemBERTa on 100,000 peptide sequences using weight-decomposed low-rank adaptation (DoRA) on a masked language model objective to generate optimized embeddings for downstream tasks involving both modified and unmodified peptides. Interestingly, when applied solely to unmodified peptides, PepDoRA did not show significant performance improvements over PeptideBERT, highlighting both the potential and limitations of alternative notational approaches.

4. Conclusion

In summary, we present a lightweight peptide language model, PepBERT with two versions. Both models achieved promising performance across a range of downstream bioactive peptide prediction tasks, establishing their potential as efficient tools for peptide representation. The datasets, source codes, pretrained models, and tutorials for usage of PepBERT are available at https://github.com/dzjxzyd/PepBERT-large. Nonetheless, more robust optimization,

particularly through hyperparameter tuning could further enhance their effectiveness. Exploring alternative notation systems for peptide language model development, such as SMILES or SELFIES, can be a worthy direction for future research. One challenge arises from the conserved structural regularity of peptide backbones. While this invariance simplifies the pattern recognition in canonical SMILES encodings, it can also introduce redundant token patterns that undermine the effectiveness of transformer-based attention mechanisms. To overcome this, a critical future direction involves designing adaptive tokenization schemes that dynamically modulate the granularity of sequential information encoding to maintain balance between sparse biological semantics (as in amino acid sequences) and chemically dense but repetitive notations like SMILES.

References:

1. Du Z, Comer J, Li Y. Bioinformatics approaches to discovering food-derived bioactive peptides: Reviews and perspectives. TrAC Trends in Analytical Chemistry 2023; 117051

2. Du Z, Li Y. Review and perspective on bioactive peptides: A roadmap for research, development, and future opportunities. Journal of Agriculture and Food Research 2022; 9:100353

3. Huang J, Xu Y, Xue Y, et al. Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences. Nat. Biomed. Eng 2023;

4. Zhang Q, Ding K, Lyv T, et al. Scientific Large Language Models: A Survey on Biological & Chemical Domains. 2024;

5. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023; 379:1123–1130

6. Thumuluri V, Martiny H-M, Almagro Armenteros JJ, et al. NetSolP: predicting protein solubility in *Escherichia coli* using language models. Bioinformatics 2022; 38:941–946

7. Xiang W, Xiong Z, Chen H, et al. FAPM: functional annotation of proteins using multimodal models beyond structural modeling. Bioinformatics 2024; 40:btae680

8. Suzek BE, Wang Y, Huang H, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2014; 31:926

9. Guntuboina C, Das A, Mollaei P, et al. PeptideBERT: A Language Model Based on Transformers for Peptide Property Prediction. J. Phys. Chem. Lett. 2023; 14:10427–10434

10. Brandes N, Ofer D, Peleg Y, et al. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 2022; 38:2102–2110

11. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019;

12. Du Z, Ding X, Xu Y, et al. UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. Briefings in Bioinformatics 2023; 1–10

13. Kumar N, Du Z, Li Y. pLM4CPPs: Protein Language Model-Based Predictor for Cell Penetrating Peptides. J. Chem. Inf. Model. 2025; 65:1128–1139

14. Ahmad W, Simon E, Chithrananda S, et al. ChemBERTa-2: Towards Chemical Foundation Models. 2022;

15. Moret M, Pachon Angona I, Cotos L, et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. Nat Commun 2023; 14:114

16. Krenn M, Häse F, Nigam A, et al. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. Mach. Learn.: Sci. Technol. 2020; 1:045024

17. Wang L, Pulugurta R, Vure P, et al. PepDoRA: A Unified Peptide Language Model via Weight-Decomposed Low-Rank Adaptation. 2024;

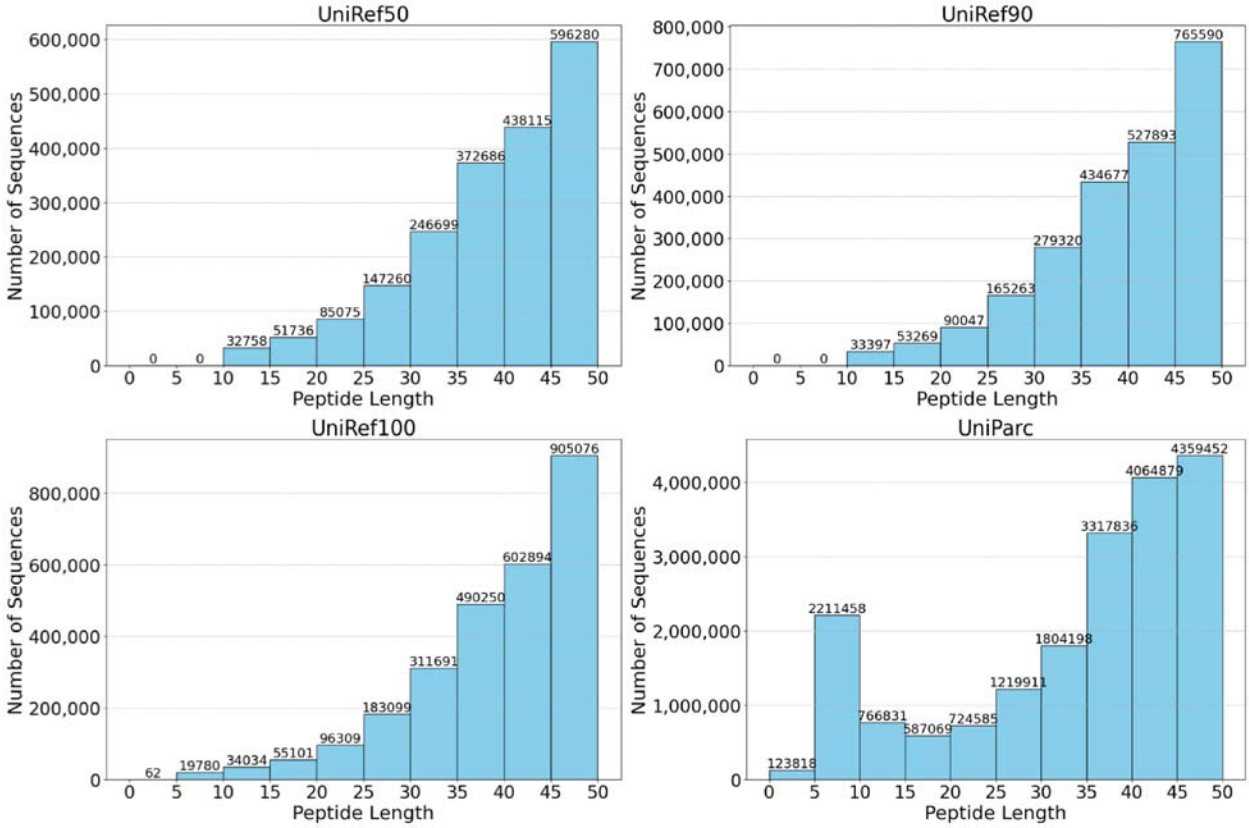

Figure 1 Peptide length distribution in each dataset for pretraining

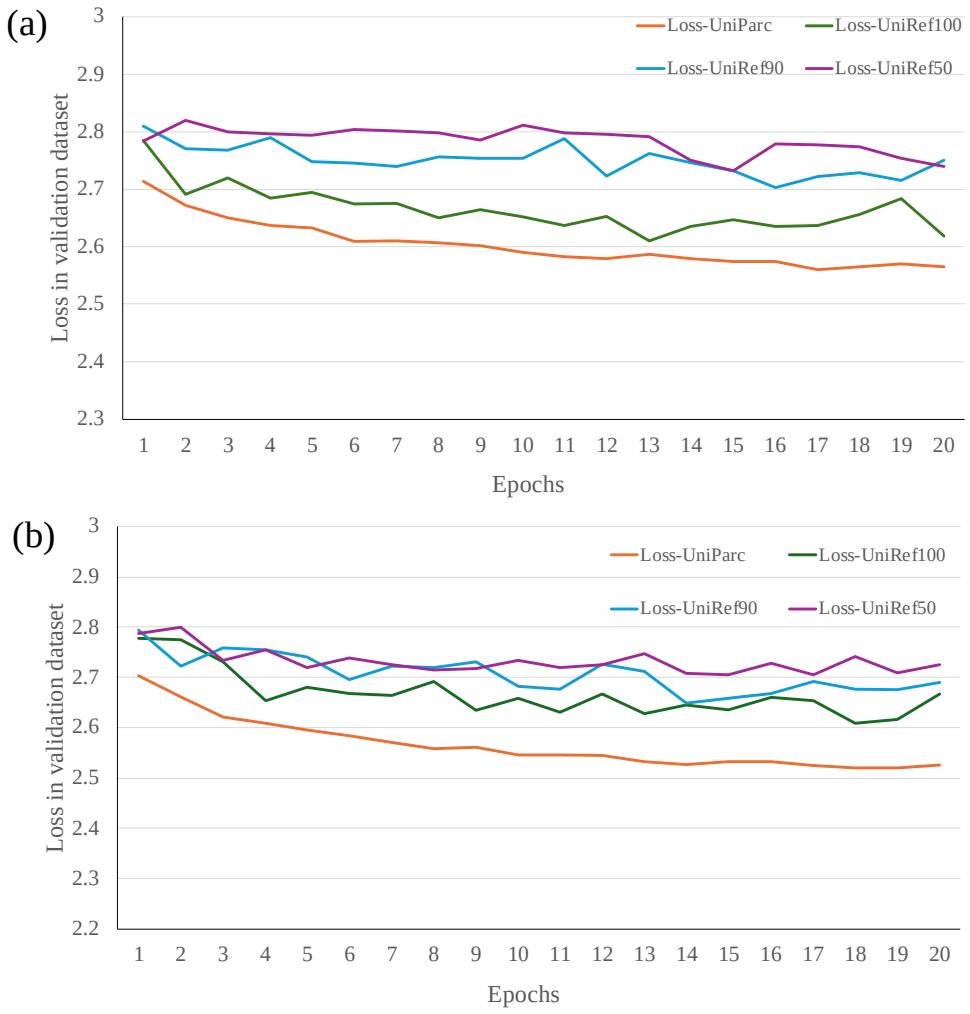

Figure 2 Loss curve of PepBERT-small (a) and PepBERT-large (b) in validation set

Table 1 Four datasets for PepBERT pretraining

Datasets	Sequence length	Number of sequences (M)	Number of amino acids (M)	Download links
UniParc	0-50	19.2	78	https://huggingface.co/datasets/dzjxzyd/UniParc_len_0_50
UniRef100	0-50	2.7	94	https://huggingface.co/datasets/dzjxzyd/UniRef100_len_0_50
UniRef90	0-50	2.35	108	https://huggingface.co/datasets/dzjxzyd/UniRef90_len_0_50
UniRef50	0-50	1.97	663	https://huggingface.co/datasets/dzjxzyd/UniRef90_len_0_50

Table 2 Comparison of PepBERT-small with ESM-2 on 9 bioactive peptide perdition tasks.

Prediction tasks	ESM-2	PepBER-small-UniParc	PepBERT-small-UniRef100	PepBERT-small-UniRef90	PepBERT-small-UniRef50
ACE inhibitory peptide	0.756±0.018	0.75±0.018**	0.747±0.02**	0.757±0.014**	0.755±0.021**
Quorum sensing peptide	0.881±0.044	0.89±0.031**	0.893±0.022**	0.885±0.028**	0.874±0.051**
Anticancer peptide-main	0.926±0.011	0.916±0.01**	0.892±0.014*	0.896±0.013*	0.901±0.012*
Anticancer peptide-alternative	0.69±0.027	0.72±0.025***	0.708±0.015**	0.699±0.019**	0.699±0.026**
Tumor T cell antigens	0.634±0.034	0.658±0.026**	0.656±0.032**	0.675±0.027***	0.65±0.028**
Blood-brain barrier penetrating peptides	0.75±0.05	0.715±0.059**	0.692±0.06*	0.746±0.081**	0.74±0.037**
Neuropeptide	0.848±0.015	0.837±0.017**	0.831±0.011*	0.822±0.015*	0.831±0.014*
Toxicity	0.883±0.014	0.883±0.008**	0.877±0.01**	0.874±0.008**	0.883±0.011**
Cell penetrating peptide	0.855±0.007	0.836±0.01*	0.836±0.01*	0.831±0.009*	0.837±0.007*

Note: performance comparison is based on balanced accuracy; ***: the performance of this model at the corresponding dataset was significantly higher than that of ESM-2 model ($p<0.05$); **: the performance of this model at the corresponding dataset had do difference with that of ESM-2 model ($p<0.05$); *: the performance of this model at the corresponding dataset was significantly lower than that of ESM-2 model ($p<0.05$).

Table 3 Comparison of PepBERT-large with ESM-2 on 9 bioactive peptide perdition tasks.

Prediction tasks	ESM-2-320	PepBERT-large-UniParc	PepBERT-large-UniRef100	PepBERT-large-UniRef90	PepBERT-large-UniRef50
ACE inhibitory peptide	0.756±0.018	0.763±0.015**	0.77±0.015**	0.763±0.014**	0.757±0.018**
Quorum sensing peptide	0.881±0.044	0.899±0.048**	0.882±0.036**	0.894±0.028**	0.884±0.037**
Anticancer peptide-alternative	0.926±0.011	0.922±0.015**	0.906±0.013*	0.913±0.011*	0.92±0.014**
Anticancer peptide-main	0.69±0.027	0.727±0.018***	0.716±0.016***	0.716±0.015***	0.711±0.028**
Tumor T cell antigens	0.634±0.034	0.672±0.027***	0.679±0.031***	0.663±0.037**	0.683±0.022***
Blood-brain barrier penetrating peptides	0.75±0.05	0.733±0.081**	0.742±0.067**	0.771±0.05**	0.725±0.044**
Neuropeptide	0.848±0.015	0.851±0.013**	0.845±0.015**	0.847±0.016**	0.852±0.015**
Toxicity	0.883±0.014	0.897±0.009***	0.902±0.01***	0.883±0.01**	0.892±0.007**
Cell penetrating peptide	0.855±0.007	0.845±0.006*	0.85±0.009**	0.861±0.007**	0.849±0.009**

Note: performance comparison is based on balanced accuracy; ***: the performance of this model at the corresponding dataset was significantly higher than that of ESM-2 model ($p<0.05$); **: the performance of this model at the corresponding dataset had do difference with that of ESM-2 model ($p<0.05$); *: the performance of this model at the corresponding dataset was significantly lower than that of ESM-2 model ($p<0.05$).