PepDoRA: A Unified Peptide Language Model via Weight-Decomposed Low-Rank Adaptation

Leyao Wang,[1,*] **Rishab Pulugurta,**[1,*] **Pranay Vure,**[1,*] **Yinuo Zhang,**[1,*] **Aastha Pal,**[1]
Pranam Chatterjee[1,2,3,†]

[1]Department of Biomedical Engineering, Duke University
[2]Department of Computer Science, Duke University
[3]Department of Biostatistics and Bioinformatics, Duke University

[*]These authors contributed equally
[†]Corresponding author: pranam.chatterjee@duke.edu

Abstract

Peptide therapeutics, including macrocycles, peptide inhibitors, and bioactive linear peptides, play a crucial role in therapeutic development due to their unique physicochemical properties. However, predicting these properties remains challenging. While structure-based models primarily focus on local interactions, language models are capable of capturing global therapeutic properties of both modified and linear peptides. Protein language models like ESM-2, though effective for natural peptides, cannot however encode chemical modifications. Conversely, pre-trained chemical language models excel in representing small molecule properties but are not optimized for peptides. To bridge this gap, we introduce PepDoRA, a unified peptide representation model. Leveraging Weight-Decomposed Low-Rank Adaptation (DoRA), PepDoRA efficiently fine-tunes the ChemBERTa-77M-MLM on a masked language model objective to generate optimized embeddings for downstream property prediction tasks involving both modified and unmodified peptides. By tuning on a diverse and experimentally valid set of 100,000 modified, bioactive, and binding peptides, we show that PepDoRA embeddings capture functional properties of input peptides, enabling the accurate prediction of membrane permeability, non-fouling and hemolysis propensity, and via contrastive learning, target protein-specific binding. Overall, by providing a unified representation for chemically and biologically diverse peptides, PepDoRA serves as a versatile tool for function and activity prediction, facilitating the development of peptide therapeutics across a broad spectrum of applications.

1 Introduction

1.1 Background

Peptide therapeutics, such as macrocycles, peptide agonists, and bioactive linear peptides—including anti-cancer, anti-microbial, anti-viral, and signaling peptides—are gaining prominence in drug development due to their unique chemical properties, such as high specificity, favorable safety profiles, and the ability to target a wide range of biological functions [1–4]. Despite their potential, predicting the functional properties of these peptides, such as bioactivity, binding affinity, and membrane permeability, remains a significant challenge [5]. This difficulty is particularly pronounced for modified peptides—those incorporating structural alterations like cyclization or unnatural amino

NeurIPS 2023 Generative AI and Biology Workshop.

acids—which exhibit therapeutic advantages over their natural counterparts, including increased stability and improved bioavailability [6, 2, 7].

Traditionally, structure-based models have been the primary tools for peptide design. Methods such as FlexPepDock, AlphaFold-based peptide co-folding, and molecular dynamics simulations are commonly used to determine the binding conformation and interactions of peptides with their targets [8–10]. However, these structure-based methods are generally not well-suited for predicting the broader therapeutic properties of peptides, such as stability, permeability, or bioactivity, as they focus primarily on local interactions rather than capturing the global sequence properties necessary for understanding the full therapeutic potential of peptides [11, 12].

Trained on millions of natural amino acid sequences, protein language models (pLMs), such as ESM-2 and ProtT5, have captured important structural and functional relationships inherent in protein sequences [13, 14]. Similar transformer-based methods have been extended to predict the properties of peptides, with models such as PeptideBERT [15]. In recent work, our lab has successfully used ESM-2 to design linear peptide binders for therapeutic applications, including models such as PepPrCLIP, PepMLM, and SaLT&PepPr, which generate effective peptide binders from sequence alone [16–18]. Without needing requirement of stable tertiary structure as input, these models enabling therapeutic design of binders and degraders to conformationally diverse targets [16–18]. Nonetheless, models like ESM-2 and PeptideBERT cannot tokenize the structural complexity of modified peptides with cyclization, non-canonical amino acids, or other modifications [13].

Contrastively, pre-trained chemical language models (cLMs), such as ChemBERTa, ChemFormer, and MolT5, are trained primarily on small molecule datasets, using SMILES notations to encode molecular structures for downstream tasks such as property prediction and molecular generation [19–21]. Because these models are largely optimized for small molecules, they do not naturally account for the sequential and structural intricacies of peptides, limiting their utility in peptide-centric applications. Recently, PeptideCLM was pre-trained on peptide-specific SMILES data from scratch, via the RoFormer model architecture, performing strongly on membrane permeabilization prediction [22]. PeptideCLM, however, was trained solely on modified peptide representations, thus the model may not retain valuable physicochemical information that pre-trained cLMs have already learned from broader molecular datasets, as well as key properties of bioactive linear peptides composed of only wild-type amino acids.

To overcome these challenges and bridge the gap between natural and modified peptide representation, we introduce PepDoRA—a unified peptide representation model designed for both bioactive linear peptides and structurally modified variants. By efficiently fine-tuning the state-of-the-art ChemBERTa-77M cLM [19] on both modified and natural peptide sequence data via efficient weight-decomposed low-rank adaptation (DoRA) [23], PepDoRA generates optimized embeddings capturing core therapeutic properties, including membrane permeability, non-fouling and hemolytic propensity, and most importantly, target-specific binding. In total, PepDoRA serves as a versatile tool for both prediction and design tasks, supporting the development of therapeutic peptides across a broad spectrum of applications.

2 Methods

2.1 Dataset

The dataset used in this study consists of 100,000 peptides, including both modified and bioactive sequences. The modified peptides were sampled from the pre-training dataset used to train PeptideCLM, incorporating cyclic structures and unnatural amino acids [22]. Bioactive linear peptides, such as anti-cancer, anti-viral, and anti-microbial peptides, were sourced from the BIOPEP-UWM database [24]. Additionally, linear peptide binders to target proteins were obtained from the Propedia v2.3 database [25]. All peptides were encoded using the SMILES string notation via RDKit [26], after which they were shuffled, combined, and then split into 80% training data and 20% testing data.

2.2 Masked Language Modeling (MLM) Task

The Masked Language Modeling (MLM) task was employed to fine-tune the ChemBERTa-77M-MLM model for peptide representation [19]. Given a peptide sequence represented as a SMILES string, a subset tokens is randomly selected and masked. The model then learns to predict the masked tokens, based on the surrounding context, to optimize the following objective:

$$L_{\text{MLM}} = -\mathbb{E}_{(x,\tilde{x})\sim D}\left[\sum_{i\in\mathcal{M}} \log P(x_i|\tilde{x})\right] \tag{1}$$

where x represents the original sequence, $\tilde{x}$ is the sequence with masked tokens, $\mathcal{M}$ denotes the set of masked positions, and $P(x_i|\tilde{x})$ is the predicted probability of the masked token x_i given the masked sequence $\tilde{x}$. The model was trained using a standard MLM approach with a masking rate of 15%.

2.3 Fine-Tuning of ChemBERTa

To adapt the ChemBERTa model for peptide prediction tasks, we employed three different fine-tuning strategies: (1) fine-tuning the last layer of ChemBERTa-77M-MLM [19], (2) using LoRA [27], and (3) using DoRA [23]. In both cases, LoRA or DoRA adapters are incorporated into the last three layers, specifically targeting the query, key, and value modules.

2.3.1 Fine-Tuning the Final Layer

In the first approach, all weights of the final transformer layer of ChemBERTa-77M-MLM [19] were unfrozen, including the attention weights, feed-forward weights, and layer normalization parameters, and updated during training. This allowed for focused adaptation of the model while minimizing the number of parameters being optimized.

2.3.2 Low-Rank Adaptation (LoRA)

LoRA aims to fine-tune a pre-trained model by adding a low-rank decomposition to the weight update [27]. Specifically, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA models the update $\Delta W \in \mathbb{R}^{d\times k}$ as:

$$W' = W_0 + BA \tag{2}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices with $r \ll \min(d, k)$. This approach allows for efficient adaptation by only updating a small number of parameters.

2.3.3 Weight-Decomposed Low-Rank Adaptation (DoRA)

DoRA decomposes the pre-trained weight into two components: magnitude (m) and direction (V), which are fine-tuned separately to enhance learning capacity [23]. The decomposition of the weight matrix W is given by:

$$W = m\frac{V}{||V||_c} \tag{3}$$

where $m \in \mathbb{R}^{1\times k}$ represents the magnitude vector, $V \in \mathbb{R}^{d\times k}$ is the directional matrix, and $||\cdot||_c$ denotes the vector-wise norm across each column of V. During fine-tuning, the weight update is given by:

$$W' = m\frac{W_0 + BA}{||W_0 + BA||_c} \tag{4}$$

Here, B and A are low-rank matrices similar to LoRA, and W_0 is the pre-trained weight. DoRA thus provides an improved adaptation mechanism while maintaining inference efficiency. The A schematic of the resultant PepDoRA is depicted in Figure 1.

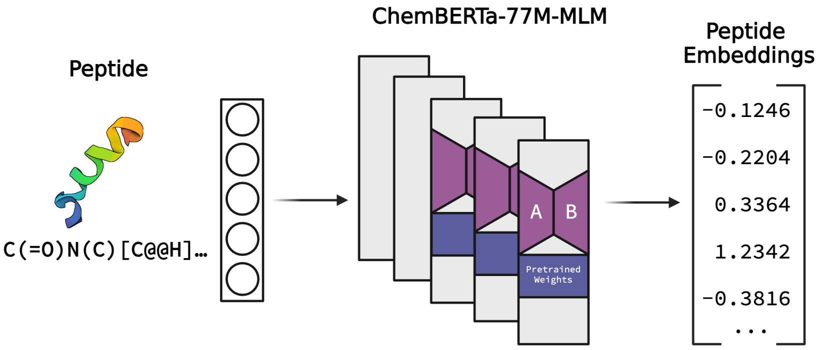

Figure 1: Schematic representation of PepDoRA for peptide embedding generation. The peptide SMILES representation is processed through the ChemBERTa-77M-MLM architecture, with the final three layers fine-tuned via DoRA, resulting in optimized peptide embeddings suitable for downstream prediction tasks.

2.4 Model Training

All models were trained on two NVIDIA A6000 GPUs, each with 48 GB of VRAM. The training of the models was conducted using the `AdamW` optimizer, with a learning rate of 2×10^{-5} and a weight decay of 0.01. The batch size for training was set to 2, and training was performed for a maximum of 10 epochs.

2.5 Evaluation

2.5.1 Membrane Permeability Prediction

The membrane permeability of peptides was predicted using a regression model identical to that described in the PeptideCLM paper [22]. Specifically, we used a regression approach to predict the logarithm of the diffusion constant for the PAMPA dataset. The dataset was split into six clusters using k-means clustering on the principal components of the peptide embeddings, with the smallest group held out for testing. The remaining data was used in a 5-fold cross-validation setup.

For evaluation, we applied this strategy across the three fine-tuned ChemBERTa models (last layer fine-tuned, LoRA, and DoRA), as well as PeptideCLM-23M and ChemBERTa-77M-MLM. For each model, the language modeling head was replaced with a fully connected feed-forward layer, with a width matching the hidden state of the model, to perform regression for predicting membrane permeability. Model training was conducted for up to 20,000 steps, and checkpoints were saved based on the lowest mean squared error (MSE) observed on the validation set. The final test metrics, including root mean squared error (RMSE), were calculated as the mean of pooled predictions from the five cross-validated models. This process was repeated five times, and the mean across the five iterations were plotted in Figure 2.

2.5.2 Non-Fouling and Hemolysis Prediction

The hemolytic activity and non-fouling property of peptides were predicted using two classification modules. Using the pre-split training datasets curated previously in [15], we utilized the XGBoost gradient boosting architecture using tree-based learners, with subsequent hyperparameter optimization via Optuna over 50 trials [28]. Model parameters were tuned by optimizing a combined metric incorporating accuracy, precision, recall and F1-score on the validation set.

We evaluated performance across the DoRA-fine-tuned ChemBERTa model, as well as PeptideBERT. PeptideBERT predictions were made using a multilayer perceptron (MLP) with a single fully connected layer of 480 nodes, trained for 30 epochs using the `AdamW` optimizer (learning rate 0.00001), batch size of 32, and binary cross-entropy loss, as described in their methods. We used the `ReduceLROnPlateau` scheduler, which reduces the learning rate by a factor of 0.1 if validation accuracy did not improve for four consecutive epochs, identical to the original implementation in [15]. Both models were evaluated using identical train-validation-test splits to ensure fair comparison.

2.5.3 Contrastive Language Model for Peptide-Protein Interaction

To evaluate peptide-protein interactions, we used a contrastive learning model based on our previous PepPrCLIP architecture [16], which itself is based on the contrastive language-image pretraining (CLIP) architecture from OpenAI [29]. The target protein embeddings were generated using ESM-2-650M, while the peptide embeddings were obtained from the three fine-tuned ChemBERTa models, PeptideCLM-23M [22], and ChemBERTa-77M-MLM [19]. The model was trained to maximize the cosine similarity between true peptide-protein pairs and minimize it for non-binding pairs:

$$L_{\text{CLIP}} = -\frac{1}{n^2}\sum_{i,j}\left[\log\frac{\exp(\text{sim}(e_i, p_j))}{\sum_{k=1}^{n}\exp(\text{sim}(e_i, p_k))}\right] \tag{5}$$

where e_i and p_j are the embeddings of the target protein and peptide, respectively, and $\text{sim}(\cdot, \cdot)$ represents the cosine similarity. Three metrics were used to evaluate the performance of this model:

- **Binary Accuracy**: This metric measures the percentage of correctly classified peptide-protein pairs, determining whether a given peptide binds to a specific protein or not.
- **Top-1 Accuracy**: This metric evaluates the model's ability to correctly identify the single best peptide-protein pair from all possible candidates. This is particularly important for identifying the most probable binding partner in cases where a direct interaction prediction is needed.
- **Top 10% Accuracy**: This metric measures whether the true peptide-protein pair is ranked within the top 10% of all possible pairs. This is particularly useful in scenarios such as high-throughput screening, where identifying a shortlist of potential candidates is of high value.

The model was trained on a curated ~12,000 peptide-protein pairs from the Propedia v2.3 [25] and PepNN [30] datasets, which were curated previously [17]. To ensure the most robust evaluation, the entire PepNN dataset was included in the training set, and the gold-standard Propedia dataset was randomly split to create separate training, validation, and test sets. Specifically, 80% of the Propedia data was allocated for training and then combined with the full PepNN training data, while the remaining 20% was split evenly into a validation set and a held-out test set, resulting in 10% for each. This splitting strategy ensured a balanced approach for model training and performance evaluation.

3 Results

Using three fine-tuned ChemBERTa variants (last layer unfrozen, LoRA, and DoRA), as well as the recent PeptideCLM-23M [22], PeptideBERT [15], and ChemBERTa-77M-MLM [19] models, we evaluated performance on predicting the therapeutic properties of peptides, including as membrane permeability, non-fouling and hemolysis propensity, and target-specific binding. These four properties are essential in assessing the therapeutic potential of peptides, particularly those with modifications, as they collectively influence a peptide's ability to enter cells, avoid nonspecific interactions, ensure biocompatibility, interact with desired targets, and achieve clinically relevant therapeutic effects.

The ability to predict membrane permeability is crucial for determining whether peptides can effectively cross cell membranes, which is a prerequisite for intracellular peptide drug targets [31, 1]. We used the PAMPA dataset [32] for evaluating membrane permeability prediction, following the benchmark strategy outlined in the PeptideCLM paper [22]. The ChemBERTa model fine-tuned via DoRA (PepDoRA) demonstrated superior performance with the lowest root mean squared error (RMSE) compared to the other models, demonstrating the model's strong adaptation capabilities to understand the chemical properties of both natural and modified peptides, making it suitable for developing therapeutic peptides capable of crossing cellular barriers (Figure 2)

Next, we compared PepDoRA to PeptideBERT on the prediction tasks of hemolysis and non-fouling, two critical properties that influence peptide safety and efficacy [15]. Hemolysis refers to the

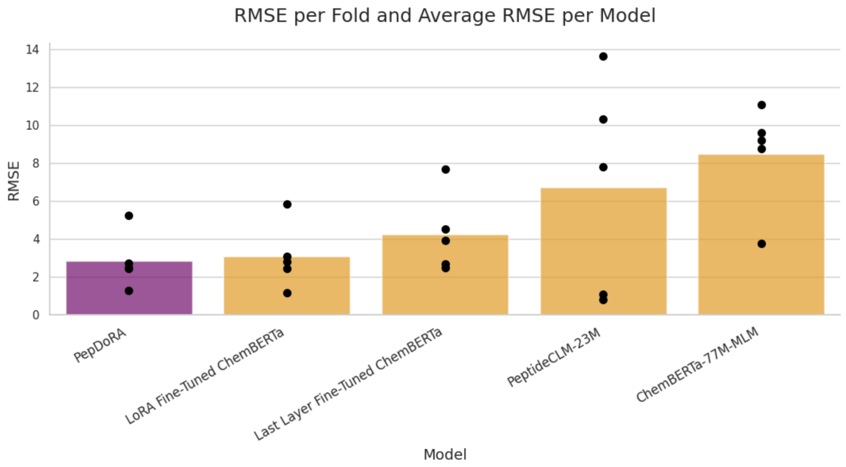

Figure 2: Comparison of membrane permeability prediction results across different model embeddings. Each bar represents the average RMSE, with individual data points indicating the RMSE values for each cross-validation fold, demonstrating the variability and consistency of each model's performance.

breakdown of red blood cells, a property essential for determining whether a peptide can be safely used without causing toxicity to blood cells [33], while non-fouling describes a peptide's resistance to nonspecific binding and surface adsorption, crucial for applications in biomaterials and drug delivery systems [34]. Both tasks were evaluated using pre-curated datasets restricted to unmodified peptides, as PeptideBERT is limited to encoding only wild-type amino acids without modifications [15]. Despite this constraint, PepDoRA demonstrated strong performance against PeptideBERT (Table 1), emphasizing its versatility to not only predict properties of modified peptides but also accurately handle linear peptide property prediction tasks.

Table 1: Evaluation of PepDoRA and PeptideBERT on non-fouling and hemolysis classification tasks using accuracy, precision, recall and F1-score metrics.

	Non-fouling		Hemolysis	
Metric	PepDoRA	PeptideBERT	PepDoRA	PeptideBERT
Accuracy	**0.87**	**0.87**	0.80	**0.81**
Precision	0.69	**0.70**	**0.59**	**0.59**
Recall	**0.71**	0.67	**0.27**	0.25
F1 Score	**0.70**	0.69	**0.37**	0.35

Accurately predicting binding interactions is crucial for designing peptides that can specifically target proteins, especially in cases involving modified peptides and undruggable targets. To assess the potential of PepDoRA embeddings for therapeutic biologics design, we evaluated its ability to capture peptide-protein interactions using a CLIP-based contrastive model. Here, we employed an architecture analogous to our recent PepPrCLIP architecture [16], using peptide embeddings in combination with target ESM-2-650M protein embeddings, with the training goal of maximizing the cosine similarity between true peptide-protein pairs (Figure 3). PepDoRA embeddings exhibited the strongest results in terms of binary accuracy, Top-1 accuracy, and Top 10% accuracy, demonstrating its ability to represent peptides effectively and enable featurization for accurate peptide-protein mapping (Table 2). Overall, these results highlight PepDoRA's utility for designing peptide therapeutics targeting specific proteins, including difficult-to-target or intrinsically disordered proteins.

4 Conclusion

In this work, we introduce PepDoRA, a new language model that fine-tunes the ChemBERTa cLM to predict the functional properties of both natural and modified peptides. By leveraging the

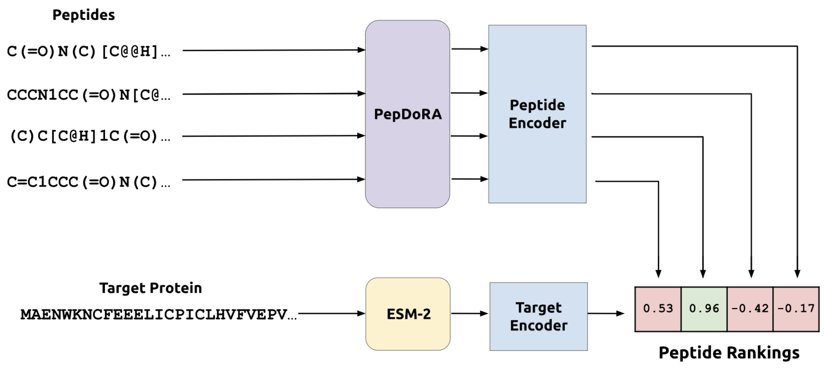

Figure 3: Schematic of the peptide-protein interaction contrastive model, illustrating learning of joint embeddings for peptide and protein pairs. The model aims to maximize cosine similarity for true binding pairs while minimizing it for non-binding pairs via a contrastive loss.

Table 2: Evaluation of models on peptide-protein interaction prediction using Binary Accuracy, Top-1 Accuracy, and Top 10% Accuracy metrics.

Model	Binary Accuracy (%)	Top-1 Accuracy (%)	Top 10% Accuracy (%)
Last Layer Fine-Tuned ChemBERTa	95.5	59.1	81.5
LoRA Fine-Tuned ChemBERTa	95.1	60.0	84.9
DoRA Fine-Tuned ChemBERTa (PepDoRA)	**95.9**	**62.1**	**86.1**
PeptideCLM-23M	93.9	47.3	74.5
ChemBERTa-77M-MLM	86.8	27.6	61.5

recently-described DoRA method for parameter efficient fine-tuning (PEFT), we demonstrate that ChemBERTa's robust physicochemical knowledge of small molecules can be effectively adapted to peptide-specific tasks without full model retraining. As a result, PepDoRA efficiently bridges the gap between chemical and peptide modeling, combining the strengths of cLMs with a PEFT strategy to generate optimized embeddings for peptides, particularly those with complex modifications like cyclization and unnatural amino acids. Our results show that PepDoRA performs strongly against alternative fine-tuning approaches, as well as the modified peptide-specific PeptideCLM model [22] and the unmodified peptide-specific PeptideBERT model [15], on multiple therapeutically-relevant tasks.

While PepDoRA's versatile representations mark substantial progress for peptide modeling, further improvements are undoubtedly possible. Scaling the model with larger, more diverse peptide datasets and leveraging additional GPU resources will enhance its current latent space. Additionally, training on richer, more diverse peptide data from high-throughput synthesis and assays may further refine the model's representation capacity. Moving forward, we plan to integrate PepDoRA-based property predictors and our target-binding contrastive model into a conditional peptide generation algorithm, using methods such as masked discrete diffusion [35], which will be followed by experimental testing of generated candidates in our lab. As PepDoRA enables peptide representation without structural information, we plan to confirm the therapeutic efficacy of these peptides for conformationally diverse disease-related targets, especially those considered undruggable by standard small molecules [36]. Ultimately, the integration of PepDoRA into our experimental workflow will advance therapeutic development, with the goal of translating these theoretical algorithms into real-world clinical solutions for previously intractable diseases.

Model Availability

PepDoRA model files and weights can be freely accessed at `https://huggingface.co/ChatterjeeLab/PepDoRA`.

References

[1] L. Wang, N. Wang, W. Zhang, X. Cheng, Z. Yan, G. Shao, X. Wang, R. Wang, and C. Fu, "Therapeutic peptides: current applications and future directions," *Signal Transduction and Targeted Therapy*, vol. 7, Feb. 2022.

[2] A. Zorzi, K. Deyle, and C. Heinis, "Cyclic peptide therapeutics: past, present and future," *Current Opinion in Chemical Biology*, vol. 38, p. 24–29, June 2017.

[3] M. A. Nauck, D. R. Quast, J. Wefers, and J. J. Meier, "Glp-1 receptor agonists in the treatment of type 2 diabetes – state-of-the-art," *Molecular Metabolism*, vol. 46, p. 101102, Apr. 2021.

[4] M. Akbarian, A. Khani, S. Eghbalpour, and V. N. Uversky, "Bioactive peptides: Synthesis, sources, applications, and proposed mechanisms of action," *International Journal of Molecular Sciences*, vol. 23, p. 1445, Jan. 2022.

[5] O. Bárcenas, C. Pintado-Grima, K. Sidorczuk, F. Teufel, H. Nielsen, S. Ventura, and M. Burdukiewicz, "The dynamic landscape of peptide activity prediction," *Computational and Structural Biotechnology Journal*, vol. 20, p. 6526–6533, 2022.

[6] Y. Li, M. Wu, Y. Fu, J. Xue, F. Yuan, T. Qu, A. N. Rissanou, Y. Wang, X. Li, and H. Hu, "Therapeutic stapled peptides: Efficacy and molecular targets," *Pharmacological Research*, vol. 203, p. 107137, May 2024.

[7] M. Oeller, R. J. D. Kang, H. L. Bolt, A. L. Gomes dos Santos, A. L. Weinmann, A. Nikitidis, P. Zlatoidsky, W. Su, W. Czechtizky, L. De Maria, P. Sormanni, and M. Vendruscolo, "Sequence-based prediction of the intrinsic solubility of peptides containing non-natural amino acids," *Nature Communications*, vol. 14, Nov. 2023.

[8] N. London, B. Raveh, E. Cohen, G. Fathi, and O. Schueler-Furman, "Rosetta flexpepdock web server—high resolution modeling of peptide–protein interactions," *Nucleic Acids Research*, vol. 39, May 2011.

[9] H. Geng, F. Chen, J. Ye, and F. Jiang, "Applications of molecular dynamics simulation in structure prediction of peptides and proteins," *Computational and Structural Biotechnology Journal*, vol. 17, p. 1162–1170, 2019.

[10] P. Bryant and A. Elofsson, "Peptide binder design with inverse folding and protein structure prediction," *Communications Chemistry*, vol. 6, Oct. 2023.

[11] H. N. Hoang, T. A. Hill, and D. P. Fairlie, "Connecting hydrophobic surfaces in cyclic peptides increases membrane permeability," *Angewandte Chemie International Edition*, vol. 60, p. 8385–8390, Mar. 2021.

[12] W.-F. Zeng, X.-X. Zhou, S. Willems, C. Ammar, M. Wahle, I. Bludau, E. Voytik, M. T. Strauss, and M. Mann, "Alphapeptdeep: a modular deep learning framework to predict peptide properties for proteomics," *Nature Communications*, vol. 13, Nov. 2022.

[13] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, p. 1123–1130, Mar. 2023.

[14] R. Schmirler, M. Heinzinger, and B. Rost, "Fine-tuning protein language models boosts predictions across diverse tasks," *Nature Communications*, vol. 15, Aug. 2024.

[15] C. Guntuboina, A. Das, P. Mollaei, S. Kim, and A. Barati Farimani, "Peptidebert: A language model based on transformers for peptide property prediction," *The Journal of Physical Chemistry Letters*, vol. 14, p. 10427–10434, Nov. 2023.

[16] S. Bhat, K. Palepu, L. Hong, J. Mao, T. Ye, R. Iyer, L. Zhao, T. Chen, S. Vincoff, R. Watson, T. Wang, D. Srijay, V. S. Kavirayuni, K. Kholina, S. Goel, P. Vure, A. J. Desphande, S. H. Soderling, M. P. DeLisa, and P. Chatterjee, "De novo design of peptide binders to conformationally diverse targets with contrastive language modeling," *bioRxiv*, June 2023.

[17] T. Chen, M. Dumas, R. Watson, S. Vincoff, C. Peng, L. Zhao, L. Hong, S. Pertsemlidis, M. Shaepers-Cheu, T. Z. Wang, D. Srijay, C. Monticello, P. Vure, R. Pulugurta, K. Kholina, S. Goel, M. P. DeLisa, R. Truant, H. C. Aguilar, and P. Chatterjee, "Pepmlm: Target sequence-conditioned generation of therapeutic peptide binders via span masked language modeling," *arXiv*, 2023.

[18] G. Brixi, T. Ye, L. Hong, T. Wang, C. Monticello, N. Lopez-Barbosa, S. Vincoff, V. Yudistyra, L. Zhao, E. Haarer, T. Chen, S. Pertsemlidis, K. Palepu, S. Bhat, J. Christopher, X. Li, T. Liu, S. Zhang, L. Petersen, M. P. DeLisa, and P. Chatterjee, "Saltnpeppr is an interface-predicting language model for designing peptide-guided protein degraders," *Communications Biology*, vol. 6, Oct. 2023.

[19] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta-2: Towards chemical foundation models," *arXiv*, 2022.

[20] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, "Chemformer: a pre-trained transformer for computational chemistry," *Machine Learning: Science and Technology*, vol. 3, p. 015022, Jan. 2022.

[21] C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho, and H. Ji, "Translation between molecules and natural language," *arXiv*, 2022.

[22] A. L. Feller and C. O. Wilke, "Peptide-specific chemical language model successfully predicts membrane diffusion of cyclic peptides," *bioRxiv*, Aug. 2024.

[23] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "Dora: Weight-decomposed low-rank adaptation," *arXiv*, 2024.

[24] P. Minkiewicz, A. Iwaniak, and M. Darewicz, "Biopep-uwm database of bioactive peptides: Current opportunities," *International Journal of Molecular Sciences*, vol. 20, p. 5978, Nov. 2019.

[25] P. Martins, D. Mariano, F. C. Carvalho, L. L. Bastos, L. Moraes, V. Paixão, and R. Cardoso de Melo-Minardi, "Propedia v2.3: A novel representation approach for the peptide-protein interaction database using graph-based structural signatures," *Frontiers in Bioinformatics*, vol. 3, Feb. 2023.

[26] A. P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L. J. Bellis, M. De Veij, and A. R. Leach, "An open source chemical structure curation pipeline using rdkit," *Journal of Cheminformatics*, vol. 12, Sept. 2020.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv*, 2021.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *arXiv*, 2019.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv*, 2021.

[30] O. Abdin, S. Nim, H. Wen, and P. M. Kim, "Pepnn: a deep attention model for the identification of peptide binding sites," *Communications Biology*, vol. 5, May 2022.

[31] N. J. Yang and M. J. Hinner, *Getting Across the Cell Membrane: An Overview for Small Molecules, Peptides, and Proteins*, p. 29–53. Springer New York, Dec. 2014.

[32] V. Siramshetty, J. Williams, D.-T. Nguyen, J. Neyra, N. Southall, E. Mathe, X. Xu, and P. Shah, "Validating adme qsar models using marketed drugs," *SLAS Discovery*, vol. 26, p. 1326–1336, Dec. 2021.

[33] A. Oddo and P. R. Hansen, *Hemolytic Activity of Antimicrobial Peptides*, p. 427–435. Springer New York, Dec. 2016.

[34] A. J. Keefe, K. B. Caldwell, A. K. Nowinski, A. D. White, A. Thakkar, and S. Jiang, "Screening nonspecific interactions of peptides without background interference," *Biomaterials*, vol. 34, p. 1871–1877, Mar. 2013.

[35] S. Goel, V. Thoutam, E. M. Marroquin, A. Gokaslan, A. Firouzbakht, S. Vincoff, V. Kuleshov, H. T. Kratochvil, and P. Chatterjee, "Memdlm: De novo membrane protein design with masked discrete diffusion protein language models," *arXiv*, 2024.

[36] X. Xie, T. Yu, X. Li, N. Zhang, L. J. Foster, C. Peng, W. Huang, and G. He, "Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials," *Signal Transduction and Targeted Therapy*, vol. 8, Sept. 2023.