Flexible and Controllable Protein Design by Prefix-tuning Large-Scale Protein Language Models

Jiawei Luo[1], Xianliang Liu[1], Jiahao Li[1], Qingcai Chen[1,2,3], Junjie Chen [1,*]

[1] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong, China;
[2] PengCheng Laboratory, Shenzhen, Guangdong, China;
[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, Shenzhen, Guangdong, China.

* Corresponding author;

E-mail addresses of the authors:
J. Luo: luojiawei@stu.hit.edu.cn
X. Liu: liuxianliang@stu.hit.edu.cn
J. Li: lijiahao@stu.hit.edu.cn
Q. Chen: qingcai.chen@hit.edu.cn
J. Chen: junjiechen@hit.edu.cn

Mail address of the corresponding author:
Junjie Chen: Room 1704, InfoTech Building, HIT Campus, Shenzhen University Town, Nanshan, Shenzhen, Guangdong, 518055, China;

Running Title: protein design with prefix-tuning protein language models

Conflict of Interest: none declared

Abstract

Designing novel proteins tailored for specific purposes presents a promising approach to address various biomedical challenges, including drug discovery, vaccine design, etc.. The protein language models (ProtLMs) with control tags as a prefix or fine-tuning in a specific domain have achieved unprecedented breakthroughs in the controllable protein design. However, the vocabulary of protein sequences only contains 20 amino acid residues, which is not like natural language vocabulary to make up flexible control tags. Moreover, due to their large number of parameters, fine-tuning ProtLMs with limited data is challenging. In this study, we propose a flexible and controllable protein design method, named PrefixProt, which employs prefix-tuning to learn a virtual token for each protein property on corresponding datasets. Subsequently, the learned virtual tokens can be used to prompt pre-trained ProtLMs for generating proteins with tailored structures and functions. We trained two prefix virtual tokens on alpha-helix structure dataset and antimicrobial peptide (AMP) dataset, respectively. Our results demonstrate that prefix virtual tokens are efficient to prompt the pre-trained ProtLM by optimizing fewer trainable parameters to achieve superior results compared with fine tuning, even under low-data settings. Furthermore, these two prefix virtual tokens are combined to precisely control protein generation with both AMP function and alpha-helix structure. These results demonstrate prefix virtual tokens are flexible to be learned and integrated to control the generation of proteins. Therefore PrefixProt has advantages of both control tags and fine-tuning. In summary, PrefixProt offers a flexible and controllable protein design solution. We anticipate that PrefixProt will contribute to protein discovery and biomedical advancement.

Keywords: controllable protein design; prefix tuning; protein language model.

Availability and implementation

The models and associated code are available at: https://github.com/chen-bioinfo/PrefixProt

1 Introduction

Proteins exhibit remarkable versatility as biomolecules, performing diverse functions in various aspects of life. Despite natural proteins' proficiency, they generally suffer from a series of limitations when repurposed for new functions: instability, not having the desired geometry or size, too low recombinant expression levels for large-scale production, exhibiting undesired interactions with other proteins or folding through too complex processes, among others [21]. The *de novo* protein design offers a solution to these challenges in therapeutics and synthetic biology [43]. In the past decades, protein engineering has achieved huge advancements through directed evolution [38]. A series of methods based on Rosetta tool [23,24] has been effectively employed to address various protein design challenges, such as enzyme design [36], vaccine design [7], and protein drug design [4]. However, this method heavily relies on homologous fragments and heuristic search methods, limiting its exploration of the vast protein sequence space to a minuscule fraction.

The vast expansion of protein sequence databases provides an opportunity for sequence-based protein design approaches. Recently, deep learning-based generative methods [13, 19, 35] are getting remarkable achievements in the field of protein design. They model the distribution underlying natural proteins in an unsupervised manner to generate novel protein variants with similar properties. For instance, the Variational Autoencoder (VAE) (MSA VAE and AR-VAE [18], Dean et al. [11], PepVAE [10]) and Generative Adversarial Networks (GANs) (ProteinGAN [35]) learn the evolutionary relationships of proteins directly from the complex amino acid sequence space, then creates new and highly diverse protein variants with natural-like physical properties. The auto-regressive models (LSTM and Transformer) also have achieved significant success in the functional protein sequence design [30,31] and signal peptide design [45]. As the pre-trained language models, such as BERT [12] and GPT [3] derived from natural language processing (NLP), were introduced into the field of protein research, protein language models (ProtLMs), such as ESM-1b [37] and ProtGPT2 [14], have been developed to generate protein sequences following the principles of natural ones.

However, above protein design methods are not controllable to design proteins with tailored functions and structures. In view of the strong ability of above methods, some researchers employed their conditional variants to disentangle protein attribute space toward controllable protein design. PepCVAE [9] employs a conditional VAE (CVAE) model to disentangle antimicrobial activity space. CLaSS [8] leverages guidance from attribute classifiers and a rejection sampling scheme for generating proteins with the desired attributes. Feedback GAN [17] using feedback mechanism to guide models to generate coding protein genes with desirable functions. To fully leverage the advantage of the pre-trained ProtLMs on protein design, some researches fine-tuned the ProtLMs on a specific domain or a particular task, allowing the model to adapt to the proteins with tailored functions. However, fine-tuning needs to modify all the model parameters. Due to the huge number of parameters of ProtLMs, adapting ProtLMs to a specific domain or a particular task, especially under data-limited scenarios, usually encounters some challenges, including overfitting, catastrophic forgetting and model drift. Inspired by the success of prompt learning [16] in NLP that utilize the prompts to control

the generation of realistic text with varied topics and sentiments, ProGen [28] was developed by taking protein property tags, such as protein family, biological process and molecular function, as prompts to control the generation of proteins. Regrettably, the vocabulary of ProtLMs is confined to 20 amino acids, which are not like natural language vocabulary to flexibly make up control tags [3] (**Figure 1A**). Nevertheless, prefix-tuning [25], a lightweight alternative to fine-tuning for generation tasks, keeps language model parameters frozen, but optimizes a small continuous task-specific vector (called the prefix). Prefix-tuning draws inspiration from prompting, allowing subsequent tokens to attend to this prefix as if it were *virtual tokens*. Thus, prefix-tuning has the advantages of both prompt learning and fine-tuning.

In this study, we propose a flexible and controllable protein design method, named PrefixProt, which introduce prefix tuning to efficiently prompt the pre-trained ProtLM for protein design with desired structures and functions. Prefix tuning is flexible to learn a virtual token for any protein property on corresponding datasets. During the training process, only the prefix virtual token is trainable, while the pre-trained ProtLM is frozen. Furthermore, multiple prefix virtual tokens can be combined to precisely control protein generation. We assess the effectiveness of PrefixProt in generating proteins with alpha-helix structure and antimicrobial function. In summary, our contributions include: (1) introducing prefix tuning to prompt the pre-trained ProtLM for protein design; (2) demonstrating the effectiveness of prefix tuning by optimizing fewer trainable parameters to achieve superior results compared with fine tuning, even under low-data settings; (3) combining two prefix virtual tokens to precisely control protein generation with both desired properties. Moreover, to the best of our knowledge, this is the first study that leverages prefix tuning to prompt ProtLMs for protein sequence generation. We aspire for this method to gain widespread adoption and contribute to the advancement of drug design and protein discovery.

2 Methods

2.1 Datasets

We constructed two protein datasets, alpha-helix structure dataset and antimicrobial function dataset, to assess the effectiveness of prefix tuning in guiding the generation of ProtGPT2 from structure and function aspects, respectively. The alpha-helix structure dataset was collected from the Alpha class of Structural Classification of Proteins (SCOP2) database [1], which is created by manual inspection and automated methods to provide a comprehensive description of the structural and evolutionary connections among known proteins. After eliminating duplication, non-protein sequences, and sequences with non-standard amino acids, the final dataset contains 9,168 proteins. The antimicrobial function dataset was collected from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP) [5, 32]. Antimicrobial peptides (AMPs) are a class of small proteins with short amino acid sequences that possess broad-spectrum antimicrobial activity for the treatment of drug-resistant bacterial infections, making them an attractive alternative to traditional antibiotics [22, 40]. AMPs were classified

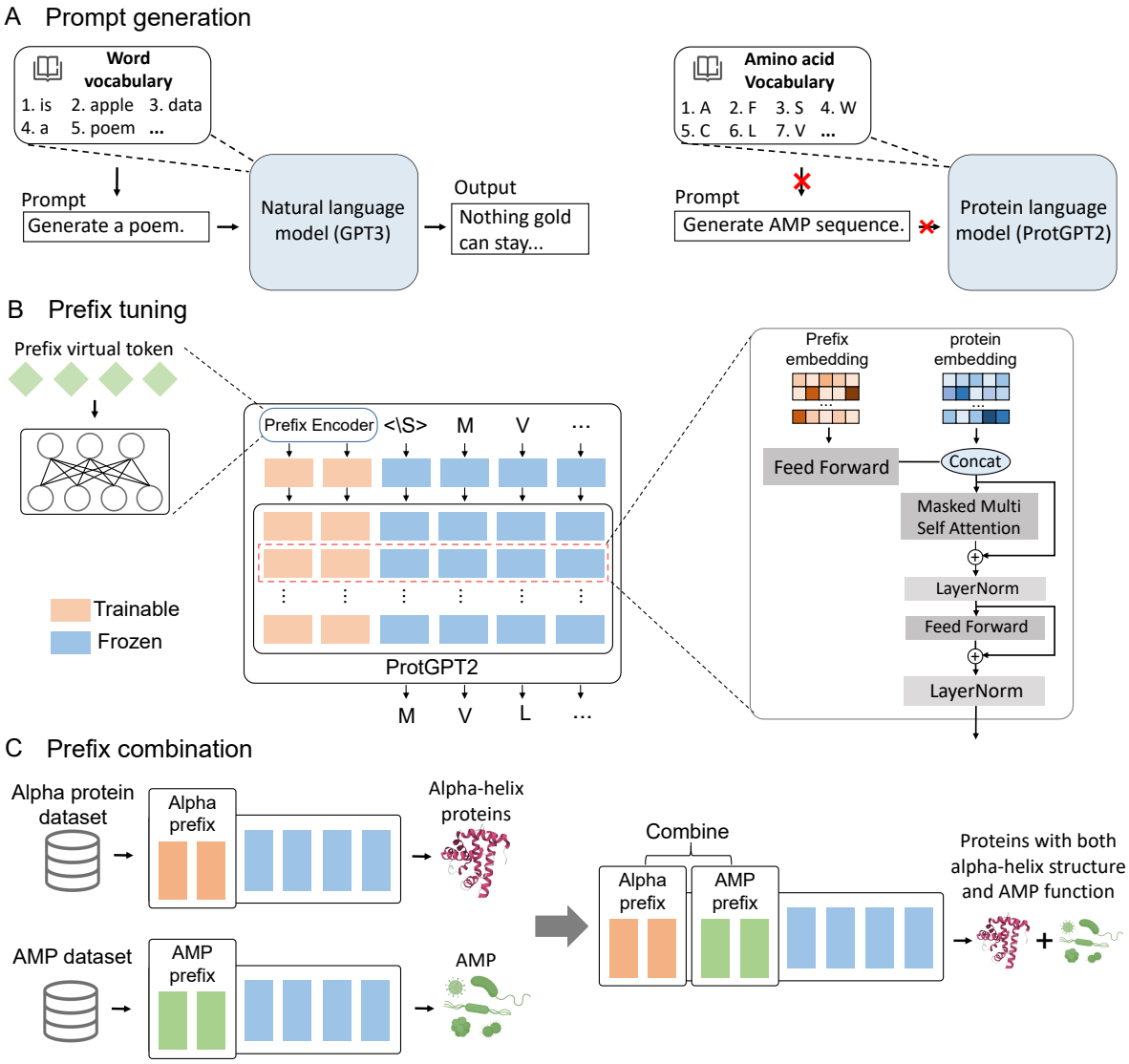

Figure 1: Illustration of the proposed method. (A) Prompt generation. The amino acid vocabulary can't be used to construct valid prompts of protein language models. (B) Prefix-tuning architecture. The prefix virtual token is concatenated on the left of the protein sequence as the first context, thereby influencing the generation of amino acids on its right. The pre-trained protein language model's parameters are frozen and only the prefix-tuning parameters are trainable. (C) Prefix combination. Prefix tuning is flexible to learn a virtual token for any protein property on corresponding datasets. Multiple prefix virtual tokens can be combined to precisely control protein generation.

as active if their IC50 against *P. aeruginosa*, *A. baumanii*, or *S. aureus* are at least one below 10 μM, 10,000 nM, or 32 μgmL^{-1}. We got a total of 4,505 active peptides. In both experiments, 80% of the data is randomly sampled as training data, and the remaining 20% is the testing data.

2.2 Autoregressive Language Model

The autoregressive language model produces the next token based on its previous context in the sequence. Let $a = (a_1, ..., a_L)$ be a protein sequence with L amino acids. The chain rule of probability can be formulated as $p(x) = \prod_{i=1}^{L} p(a_i|a_{<i})$ [2]. An autoregressive ProtLM is to learn the distribution of amino acids in proteins, called linkage disequilibrium or evolutionary patterns in Proteomics. In practically, a ProtLM with parameters θ can be trained by minimizing the negative log-likelihood loss value across a given protein dataset $D = \{a^1, ..., a^{|D|}\}$. The loss function can be formulated as

$$L = \sum_{k=1}^{|D|} logp_\theta(a_i^k|a_{<i}^k) \ . \tag{1}$$

Here, we employed a pre-trained language model ProtGPT2 [14] as the base model for protein generation. ProtGPT2 is a transformer decoder-only language model trained on UniRef50. Its architecture is similar to the previously released GPT2-large Transformer [33], which is constructed with a 36-layer transformer decoder and 12 attention heads in each layer with a model dimensionality of 1280.

2.3 Prefix-Tuning

Prefix tuning trains a virtual token as a prompt to guide ProtLM in generating proteins with desired functions and structures. The prefix token is a virtual token because it does not represent specific amino acids but is learned through training on a specialized dataset. The prefix token p is concatenated on the left of the protein sequence as the first context, thereby influencing the generation of amino acids on its right. Let $a' = [p; a]$ denote the concatenated input, p_i and $|p|$ be the i-th virtual prefix token and length of prefix respectively. The prefix embedding is generated by the prefix encoder $P_\beta \in R^{|p| \times d}$. The output h_i of the ProtLM with a prefix can be formulated as follows:

$$h_i = \begin{cases} P_\beta(i), \text{if } i < |P| \\ ProtLM_\sigma(a_i', h_{<i}), otherwise \end{cases} , \tag{2}$$

where σ, β are parameters of the ProtLM and prefix encoder. The objective function of prefix tuning is the same as the autoregressive language model. Furthermore, to deal with the unstable optimization of parameters β, $P_\beta[i, :]$ is reparameterized by a matrix (P'_β) into $MLP_\beta(P'_\beta[i, :])$ by using feedforward neural network. During training, the parameters σ in the ProtLM are frozen, while only the parameters β in the prefix encoder are trainable. Thus, after training only the prefix encoder P_β needs to be stored.

2.4 Sequence Generation

During the generation process, prefix tokens associated with desired structures and functions are first fed into prefix-tuned ProtLM. Then in every generation step, a subsequent

amino acid is produced according to its previous tokens. To increase the generative diversity and reduce the repeated amino acids, top-k sampling [33] with a repetition penalty [20] is employed, which is a sampling strategy that selects the subsequent tokens from the top k most probable tokens featured within the distribution generated by the ProtLM. Typically, the top k tokens are sampled by using the softmax function with a sampling temperature. The probability of the i-th token being sampled from the top k tokens is computed by Eq. (3).

$$p_i = \frac{\exp(\mathrm{z_i}/\mathrm{T})}{\sum_j \exp(\mathrm{z_j}/\mathrm{T})} \tag{3}$$

where z_i is the prediction for token i, T is the temperature, and p_i is the sampling probability of token i. The next token is then chosen by sampling through a multinomial distribution with probabilities p_i clipped at the top-k tokens. To prevent repetition, a repetition penalty is used in top-k sampling to reduce the probabilities of tokens that have been presented within priors. The probability of the i-th token is defined as Eq.(4)

$$p_i = \frac{\exp(z_i/T \cdot I(i \in g))}{\sum_j \exp(z_j/T \cdot I(j \in g)}, \ I(i \in g) = \theta \text{ if } i \in g \text{ is True else 1}, \tag{4}$$

where g represents a list of generated tokens.

2.5 Evaluation Details

We measured the quality of generated protein sequences from structure and function aspects by utilizing several metrics, including perplexity, percentage of alpha-helix, Rosetta energy score, probability of antimicrobial function, as well as the identity and diversity.

Perplexity is a general metric in sequence generation tasks, which is the exponential value of the cross-entropy loss computed for each token within a given dataset. A lower perplexity correlates with a higher quality model. The percentage of alpha-helix was calculated by employing the DSSP tool to count the amino acids with alpha-helix structure in the protein structure predicted by ESMFold [26]. We got Rosetta energy scores by running Rosetta Relax with the default Rosetta full-atom energy terms and weights in the Pyrosetta [6]. Lower Rosetta energy conformers correlate with more relaxed structures [39]. To minimize the time needed for relax, we set the maximum number of iterations to 20. A general guideline is that the total score (Rosetta Energy Units, REU) should typically fall within the range of -1 to -3 per residue [44]. We calculated the physicochemical properties of generated AMPs, including length, charge, charge density, hydrophobicity, hydrophobic moment, and isoelectric point by using modlAMP [29]. The antimicrobial function of generated AMPs were predicted by leveraging the CAMP server [42]. Sequences with a length greater than 100 amino acids were removed before prediction. The sequence identity was performed by BLAST.

2.6 Implementation details

We employed the pre-trained ProtGPT2 as the base model, and optimized it with prefix tuning and fine tuning methods on alpha-helix structure dataset and antimicrobial

function dataset. To estimate the effectiveness of prefix tuning, we compared the proteins generated by prefix-tuned ProtGPT2 with the proteins generated by pre-trained ProtGPT2 and fine-tuned ProtGPT2, as well as two baselines: the natural proteins in testing data and the randomly generated proteins.

All compared methods were trained in 100 epochs with the AdamW optimizer [27]. We set a learning rate of 5e-6 and 5e-3 for fine-tuning and prefix tuning with batch size 8 on the alpha-helix structure dataset, and a learning rate of 1e-6 and 1e-3 for fine-tuning and prefix tuning with batch size 32 on antimicrobial function dataset, respectively. We implemented a polynomial decay schedule with a final learning rate of 1e-7 and a polynomial exponent of 3. The length of prefix virtual tokens is 100 and 20 for the alpha-helix structure dataset and antimicrobial function dataset, respectively. During generation, the subsequent amino acids were sampled from the top 500 with repetition penalty of 1.2. The random dataset was created by randomly combining 20 amino acids according to their background distribution in the training dataset. We generated 2,000 proteins with alpha-helix structure and 1000 proteins with antimicrobial function by each compared method.

3 RESULT

3.1 Controllable alpha-helix protein generation

In this section, we employed prefix-tuning to learn a virtual token on the alpha-helix structure dataset, and evaluated the effectiveness of prefix tuning to prompt ProtGPT2 from the structure aspect. The prefix-tuned ProtGPT2 was compared with the pre-trained ProtGPT2, fine-tuned ProtGPT2, as well as random generation according to amino acid background distribution.

We first compared the percentage of amino acids with alpha-helical structure and the Rosetta energy score (Rosetta Energy Units, REU) on the generated proteins by all compared methods against the testing dataset. Prefix-tuned ProtGPT2 outperforms other methods in terms of both metrics. For the alpha-helix structures comparison (**Figure 2A**), although both prefix-tuning and fine-tuning methods significantly improve the percentage of alpha-helix structure in generated proteins comparing to pre-trained ProtGPT2, the prefix tuning achieves closer distribution to natural proteins in the testing dataset, having a higher proportion of amino acids with alpha-helix structures. To further estimate the model's capability to efficiently generate stable proteins, we calculated the Rosetta energy of generated proteins (**Figure 2B**). The energy scores of majority proteins are between -1 and -3 REU/residue, except proteins in the random generation dataset. The mean energy scores for testing, prefix-tuned, fine-tuned datasets are -1.59, -0.92, and -0.8 REU/residue, respectively. The proteins generated by prefix-tuned ProtGPT2 have lower Rosetta energy scores than the proteins generated by the other methods, but are still slightly higher than the natural proteins in the testing dataset. The three 'all alpha' protein examples (**Figure 2C**) generated by prefix-tuned ProtGPT also show the stability with low energy scores.

To further analyze the conformation and dihedral angular distribution of proteins,

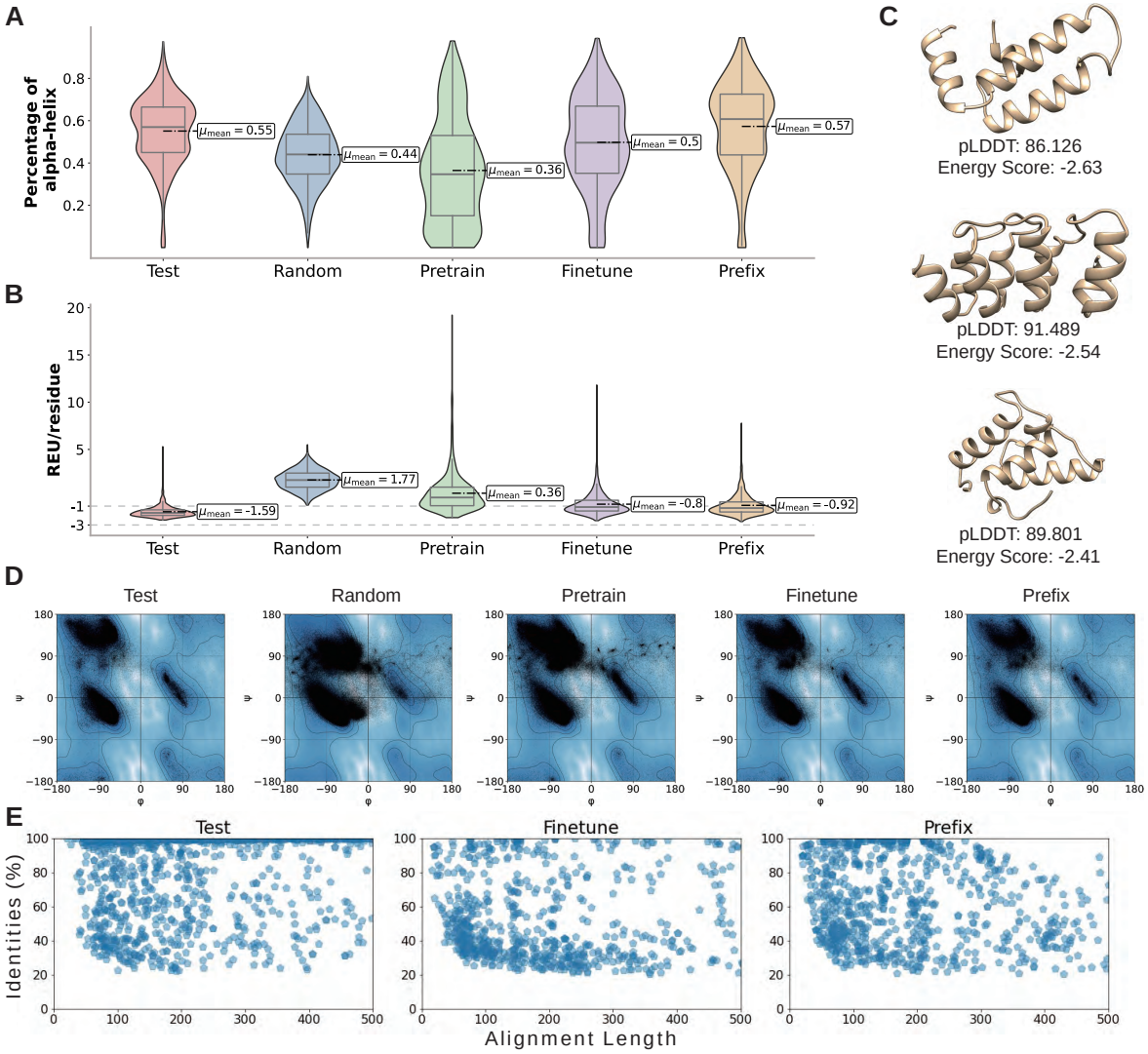

Figure 2: Comparison in generating proteins with alpha-helix structure. (A) Percentage of amino acids within alpha-helix structure in the compared datasets. (B) Average Rosetta energy units per residue in the compared datasets. (C) Three proteins with 'all alpha-helix' generated by prefix-tuned ProtGPT2. (D) Ramachandran plots comparing the (ϕ, ψ) dihedral angles for the testing set and generated proteins by different methods. (E) Pairwise sequence identity comparison of testing dataset, proteins generated by fine-tuned and prefix-tuned ProtGPT2 against training dataset.

we conducted Ramachandran plots [34] to illustrate the co-occurrence frequency of dihedrals (ϕ, ψ) (**Figure 2D**). The results indicate that all methods are able to generate all three major secondary structure elements in protein backbones. Nonetheless, the proteins generated by prefix tuning exhibit greater similarity to the natural sequence conformation, while other methods produce dihedral angles that are more dispersed and less reasonable, scattered across the periphery. We also conducted a comparison of sequence identity for generated proteins against the training dataset (**Figure2E**). The

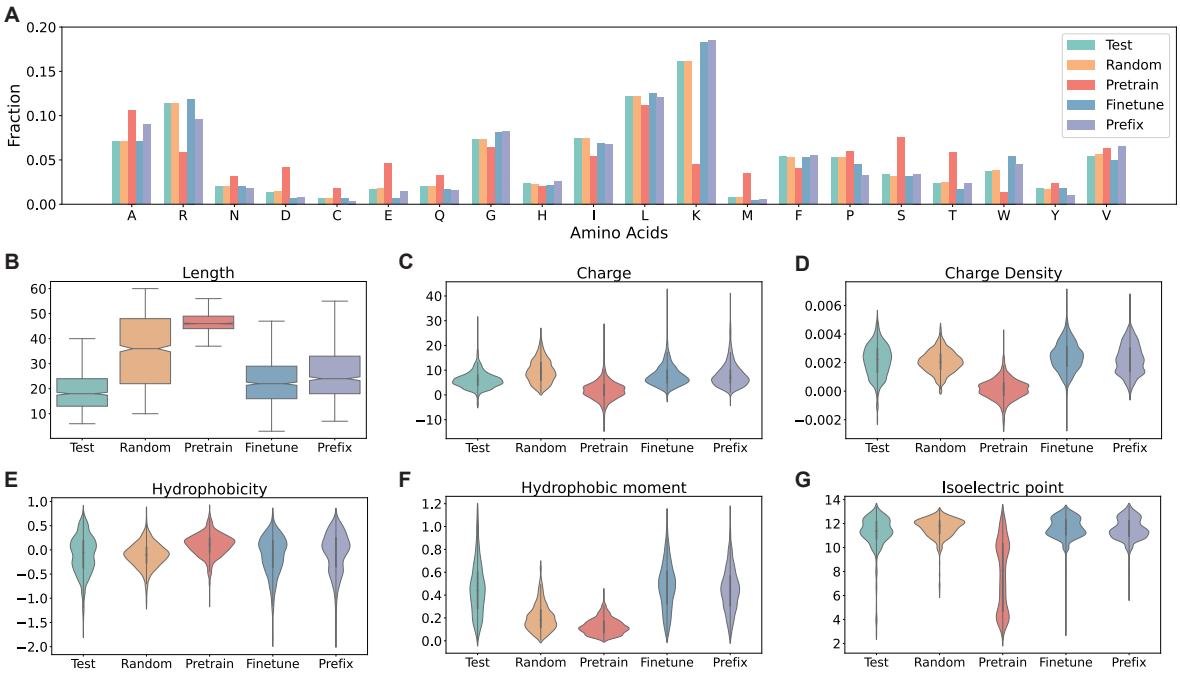

Figure 3: Physicochemical property comparison in generating proteins with antimicrobial function.

proteins produced through prefix tuning exhibit a higher sequence identity, whereas the proteins generated by fine-tuning predominantly display identities falling within the range of 20% to 50%. Note that the proteins generated by pre-trained ProtGPT2 and random generation have few sequences matched with training data. Thus their results are not presented.

Overall, prefix tuning with a limited number of trainable parameters surpasses fine tuning with fully trainable parameters in generating proteins with stable energy and reasonable conformation, demonstrating the capability of prefix tuning to prompt large ProtLM with desired structures.

3.2 Controllable antimicrobial protein generation

In this section, we trained a prefix virtual token on an antimicrobial protein dataset, and evaluated the capability of prefix tuning from the antimicrobial function aspect. We first compared the generated proteins in terms of 7 physicochemical properties, including amino acid distribution, length, charge, charge density, hydrophobicity, hydrophobic moment, and isoelectric point (**Figure 3**). Since the ProtGPT2 was pre-trained on a large-scale protein dataset, which includes peptides with short amino acid sequences and large proteins with long amino acid sequences, we limited the maximum length to 20 amino acids in the generation process to focus on peptide generation. Except for AMPs generated by pre-trained ProtGPT2, all other methods are able to produce AMPs with amino acid distribution close to natural AMPs in the testing dataset (**Figure 3**A). Note that the random generation is based on the background distribution of amino

acids in the testing dataset. Thus it has the same distribution with testing data. The other results of 6 physicochemical properties distribution show that the AMPs generated by prefix tuning and fine tuning are both notable resemblance in physicochemical properties of natural AMPs in the testing dataset, demonstrating that both of them are potential to direct ProtGPT2 toward the generation of protein sequences with target physicochemical properties [15].

We then compared antimicrobial activity of generated AMPs between prefix tuning and several state-of-the-art AMP generation methods, including PepLSTM [30], PepCVAE [9], and Hydramp [41]. PepLSTM designs AMPs by leveraging Long Short-Term Memory (LSTM) to recognize the grammar of amphipathicity in peptides. PepCVAE designs novel AMPs based on a semi-supervised variational autoencoder model. HydrAMP employed a conditional variational autoencoder to project amino acid sequences into reduced-dimensional, continuous representations. The activity of all generated AMPs was predicted by CAMP [42], an open-access AMP prediction web-server (**Figure 4A**).

It is obvious that the prefix-tuned ProtGPT2 produces AMPs with higher probability of being active than the fine-tuned ProtGPT2. And the activity distributions of AMPs generated by prefix-tuned ProtGPT2 and fine-tuned ProtGPT2 are significantly higher than other methods. Similarly, we also conducted a comparison of sequence identity for generated AMPs against the training dataset (**Figure 4B**). AMPs produced by prefix tuning exhibit similar distribution to natural AMPs, demonstrating they have higher sequence identity. In contrast, only one sequence generated by fine-tuning can be matched to the training set. Furthermore, the amphipathicity has been demonstrated to be a significant property influencing the antimicrobial activity of antimicrobial peptides [15]. We generated helical wheel plots of 5 top-ranking AMPs generated by the prefix-tuned ProtGPT2 (**Figure 4C**). The plots show that the generated AMPs have the formation of amphipathic helical structures. The results suggest that prefix tuning can prompt ProtGPT2 to generate proteins with desired functions.

3.3 Combing prefix tokens for fine-grained control in protein generation

A large language model can be prompted by a complicated instruction that contains multiple tokens for fine-grained control, such as using multiple tokens to specify different aspects or conditions within the instruction. To further investigate the potential of prefix tuning in protein generation, we combined alpha-helix prefix token and AMP prefix token to prompt the ProtGPT2 to generate proteins with both alpha-helix structure and AMP function.

We implemented two manners for combining the alpha-helix prefix token and AMP prefix token: concatenating or averaging their prefix embeddings (**Figure 5**). Regarding the limitation of input length and robustness of token combination, we retrained the alpha-helix prefix embedding with length of 20. To illustrate the impact of the AMP prefix token, we integrated the AMP prefix token into the alpha-helix prefix token, and conducted an ablation experiment by replacing the AMP prefix token with randomly initial prefix tokens. Since the AMPs are a class of small proteins with short

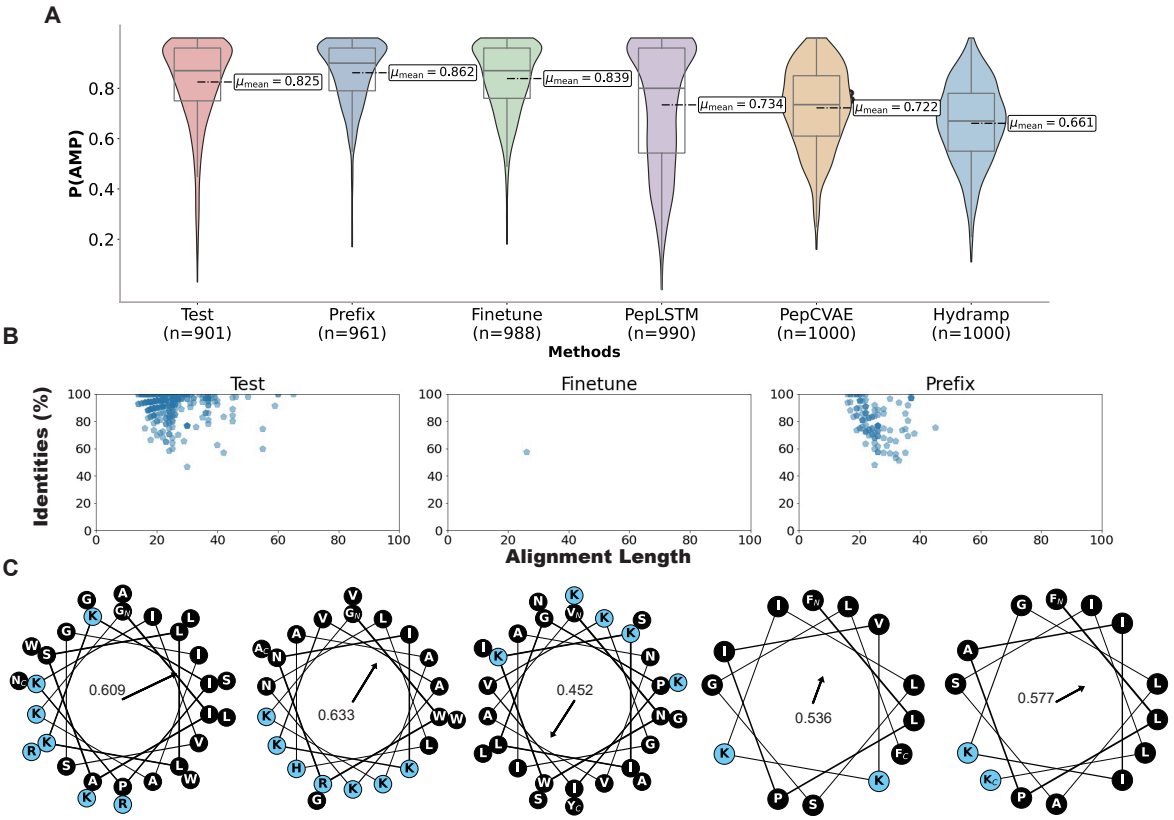

Figure 4: Antimicrobial activity comparison in generating proteins with antimicrobial functions. (A) The probability distribution comparison of being the active AMPs. (B) Pairwise sequence identity comparison of testing dataset, generated by fine-tuned and prefix-tuned ProtGPT2 against training dataset. (C) The helical wheel plots of 5 top-ranking AMps generated by the prefix-tuned ProtGPT2. The amino acids labeled with "N" and "C" subscript represent the N-terminus and C-terminus of the peptide sequences, respectively. The black circles represent hydrophobic residues, while the blue circles represent polar residues. The spatial segregation between hydrophobic and polar amino acids can be precisely assessed using the hydrophobic moment (HM) and the associated HM vector points toward the hydrophobic face of the helix.

amino acid sequences, we filtered out the generated sequences with length exceeding 100 amino acids. After filtering, the peptides prompted by only alpha-helix prefix token are left with 254 sequences, with most of them having length around 80 amino acids and the average activity probability of 0.425. When combining the AMP prefix token into the alpha-helix prefix token, especially using the averaging manner, the probability distribution of being AMP is notably improved and the length of generated peptides becomes shorter. However, the randomly initial prefix token has no discernible impact on the peptide generation. These results demonstrate that combining multiple prefix tokens can effectively fine-grain the protein generation by specifying different properties or conditions. Taking this strength, we can construct more prefix tokens to guide the ProtLM's behavior in a more precise manner, leading to more accurate and desired outputs.

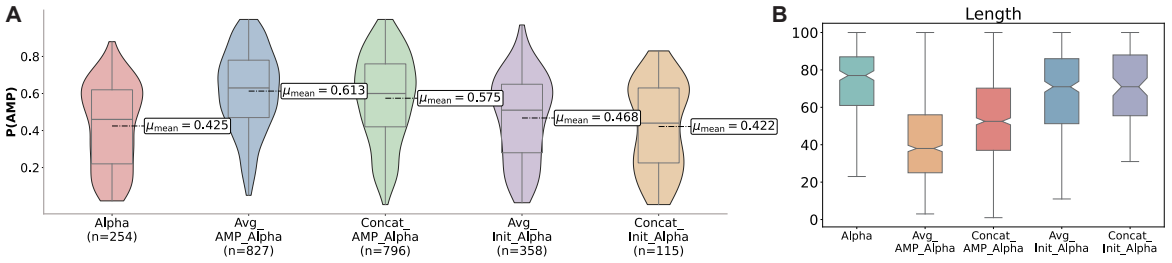

Figure 5: Comparison of probability of antimicrobial activity (A) and sequence length (B) in generating proteins with antimicrobial functions by using combined prefix tokens.

3.4 Impact of low training data and length of prefix embedding

To evaluate the effectiveness of prefix tuning in protein generation tasks in low-data scenarios, we randomly subsampled the full antimicrobial function dataset to obtain 6 small datasets with 50, 100, 200, 500, 1000, 2000 AMPs. We compared fine-tuned and prefix-tuned ProtGPT2 in a low-data setting (**Figure 6A, B**). Although both of them exhibit an improvement trend in terms of activity probability of generated AMPs as the size of the training data increases, the prefix tuning outperforms the fine-tuning in all instances of low-data settings, especially on the 4 lower-data settings with 50, 100, 200, 500 AMPs. But the gap narrows down as the dataset size increases. These results highlight the effectiveness of prefix tuning when dealing with small datasets.

A longer prefix embedding results in more trainable parameters, leading to increased expressive power in the prompt generation, but more parameters also need larger data to train. We evaluated the prefix tuning with different prefix embedding length settings (**Figure 6C-E**). In the alpha-helix protein generation, better Rosetta energy scores and comparable alpha-helix distributions are achieved when the prefix length is 100. The performance reaches the peak at a threshold of 20 in the AMP design task and then slightly drops as the length increases. The prefix length is indeed a crucial hyperparameter in prefix tuning.

4 Conclusion

In this study, we introduced the prefix tuning to efficiently prompt large protein language models for generating proteins with desired properties. We evaluated the effectiveness of prefix tuning on generating proteins with alpha-helix structure and with antimicrobial function. The results indicate prefix tuning is efficient than fine tuning in terms of higher performance and fewer trainable parameters. Moreover, multiple prefix tokens can be combined to guide the protein language model in a more precise manner, leading to more accurate and desired protein design.

In future study, we plan to integrate multiple modal information including structure-related representation and function-related features for protein generation. These methods can be effectively extended to various applications in the field of antibody and drug peptide discovery, thereby facilitating advancements in drug discovery research.

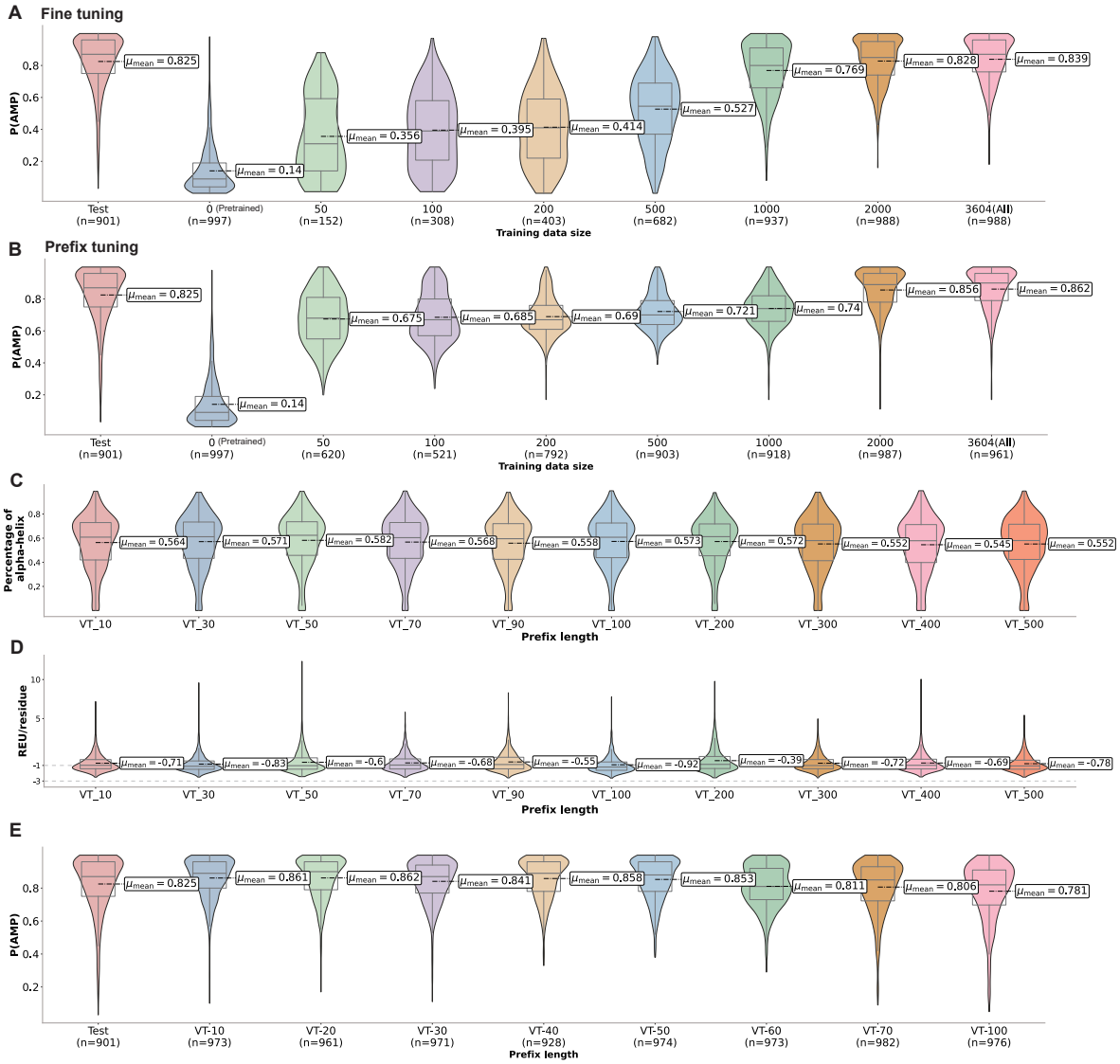

Figure 6: Comparison of antimicrobial activity in low-data settings and various prefix length settings. The probability distribution for AMPs generated by the fine-tuned ProtGPT2 (A) and the prefix-tuned ProtGPT2 (B) in a low-data setting. (C and D) The percentage of amino acids within alpha-helix structure and average Rosetta energy units per residue in the datasets generated by the prefix-tuned ProtGPT2 with various prefix length settings. (E) The probability distribution for AMPs generated by the prefix-tuned ProtGPT2 with various prefix length settings.

5 Acknowledgements

This work was supported by the National Key R&D Program of China (2022ZD0116002), National Natural Science Foundation of China (62102118, 62276075, 61872113), Project of Educational Commission of Guangdong Province of China (2021KQNCX274), Shenzhen Colleges and Universities Stable Support Program (GXWD20220811170504001),

Shenzhen Science and Technology Program (JCYJ20190806112210067), Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, China (2022B121 2010005).

References

[1] Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, and Alexey G Murzin. Scop2 prototype: a new approach to protein structure mining. *Nucleic acids research*, 42(D1):D310–D314, 2014.

[2] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022.

[5] Alice Capecchi, Xingguang Cai, Hippolyte Personne, Thilo Köhler, Christian van Delden, and Jean-Louis Reymond. Machine learning designs non-hemolytic antimicrobial peptides. *Chemical science*, 12(26):9221–9232, 2021.

[6] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics*, 26(5):689–691, 2010.

[7] Bruno E Correia, John T Bates, Rebecca J Loomis, Gretchen Baneyx, Chris Carrico, Joseph G Jardine, Peter Rupert, Colin Correnti, Oleksandr Kalyuzhniy, Vinayak Vittal, et al. Proof of principle for epitope-focused vaccine design. *Nature*, 507(7491):201–206, 2014.

[8] Payel Das, Tom Sercu, Kahini Wadhawan, Inkit Padhi, Sebastian Gehrmann, Flaviu S. Cipcigan, Vijil Chenthamarakshan, Hendrik Strobelt, Cicero dos Santos, Pin-Yu Chen, Yi Yan Yang, Jeremy P. K. Tan, James L. Hedrick, Jason Crain, and Aleksandra Mojsilovic. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, 5:613 – 623, 2021.

[9] Payel Das, Kahini Wadhawan, Oscar Chang, Tom Sercu, Cicero Dos Santos, Matthew Riemer, Vijil Chenthamarakshan, Inkit Padhi, and Aleksandra Mojsilovic. Pepcvae: Semi-supervised targeted design of antimicrobial peptide sequences. *arXiv preprint arXiv:1810.07743*, 2018.

[10] Scott N. Dean, Jerome Anthony E. Alvarez, Dan Zabetakis, Scott A. Walper, and Anthony P. Malanoski. Pepvae: Variational autoencoder framework for antimicrobial peptide generation and activity prediction. *Frontiers in Microbiology*, 12, 2021.

[11] Scott N Dean and Scott A Walper. Variational autoencoder for generation of antimicrobial peptides. *ACS omega*, 5(33):20746–20754, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Wenze Ding, Kenta Nakai, and Haipeng Gong. Protein design via deep learning. *Briefings in bioinformatics*, 23(3):bbac102, 2022.

[14] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[15] Christopher D Fjell, Jan A Hiss, Robert EW Hancock, and Gisbert Schneider. Designing antimicrobial peptides: form follows function. *Nature reviews Drug discovery*, 11(1):37–51, 2012.

[16] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

[17] Anvita Gupta and James Zou. Feedback gan for dna optimizes protein functions. *Nature Machine Intelligence*, 1(2):105–111, 2019.

[18] Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLoS computational biology*, 17(2):e1008736, 2021.

[19] Michael Jendrusch, Jan O Korbel, and S Kashif Sadiq. Alphadesign: A de novo protein design framework based on alphafold. *Biorxiv*, pages 2021–10, 2021.

[20] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[21] Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, and David Baker. Principles for designing ideal protein structures. *Nature*, 491:222–227, 2012.

[22] Brian P Lazzaro, Michael Zasloff, and Jens Rolff. Antimicrobial peptides: Application informed by evolution. *Science*, 368(6490):eaau5480, 2020.

[23] Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian W Kaufman, P Douglas Renfrew, Colin A Smith,

Will Sheffler, et al. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. In *Methods in enzymology*, volume 487, pages 545–574. Elsevier, 2011.

[24] Julia Koehler Leman, Brian D Weitzner, Steven M Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F Alford, Melanie Aprahamian, David Baker, Kyle A Barlow, Patrick Barth, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, 17(7):665–680, 2020.

[25] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[26] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[28] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

[29] Alex T Müller, Gisela Gabernet, Jan A Hiss, and Gisbert Schneider. modlamp: Python for antimicrobial peptides. *Bioinformatics*, 33(17):2753–2755, 2017.

[30] Alex T Muller, Jan A Hiss, and Gisbert Schneider. Recurrent neural network model for constructive peptide design. *Journal of chemical information and modeling*, 58(2):472–479, 2018.

[31] Deepesh Nagarajan, Tushar Nagarajan, Natasha Roy, Omkar Kulkarni, Sathyabaarathi Ravichandran, Madhulika Mishra, Dipshikha Chakravortty, and Nagasuma Chandra. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *Journal of Biological Chemistry*, 293(10):3492–3509, 2018.

[32] Malak Pirtskhalava, Anthony A Amstrong, Maia Grigolava, Mindia Chubinidze, Evgenia Alimbarashvili, Boris Vishnepolsky, Andrei Gabrielian, Alex Rosenthal, Darrell E Hurt, and Michael Tartakovsky. Dbaasp v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic acids research*, 49(D1):D288–D297, 2021.

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[34] GN t Ramachandran and V Sasisekharan. Conformation of polypeptides and proteins. *Advances in protein chemistry*, 23:283–437, 1968.

[35] Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3(4):324–333, 2021.

[36] Florian Richter, Andrew Leaver-Fay, Sagar D Khare, Sinisa Bjelic, and David Baker. De novo enzyme design using rosetta3. *PloS one*, 6(5):e19230, 2011.

[37] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[38] Philip A Romero and Frances H Arnold. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology*, 10(12):866–876, 2009.

[39] Marion F Sauer, Alexander M Sevy, James E Crowe Jr, and Jens Meiler. Multi-state design of flexible proteins predicts sequences optimal for conformational change. *PLOS Computational Biology*, 16(2):e1007339, 2020.

[40] Josep M Sierra, Ester Fusté, Francesc Rabanal, Teresa Vinuesa, and Miguel Viñas. An overview of antimicrobial peptides and the latest advances in their development. *Expert opinion on biological therapy*, 17(6):663–676, 2017.

[41] Paulina Szymczak, Marcin Możejko, Tomasz Grzegorzek, Radosław Jurczak, Marta Bauer, Damian Neubauer, Karol Sikora, Michał Michalski, Jacek Sroka, Piotr Setny, et al. Discovering highly potent antimicrobial peptides with deep generative model hydramp. *Nature Communications*, 14(1):1453, 2023.

[42] Faiza Hanif Waghu, Lijin Gopi, Ram Shankar Barai, Pranay Ramteke, Bilal Nizami, and Susan Idicula-Thomas. Camp: Collection of sequences and structures of antimicrobial peptides. *Nucleic acids research*, 42(D1):D1154–D1158, 2014.

[43] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, T. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1089 – 1100, 2023.

[44] Michael J Wedemeyer, Benjamin K Mueller, Brian J Bender, Jens Meiler, and Brian F Volkman. Modeling the complete chemokine–receptor interaction. In *Methods in cell biology*, volume 149, pages 289–314. Elsevier, 2019.

[45] Zachary Wu, Kevin K Yang, Michael J Liszka, Alycia Lee, Alina Batzilla, David Wernick, David P Weiner, and Frances H Arnold. Signal peptides generated by attention-based neural networks. *ACS Synthetic Biology*, 9(8):2154–2161, 2020.