Protein Language Model Supervised Precise and Efficient Protein Backbone Design Method

Bo Zhang[#], Kexin Liu[#], Zhuoqi Zheng, Yunfeiyang Liu, Junxi Mu, Ting Wei*, Hai-Feng Chen*

State Key Laboratory of Microbial metabolism, Joint International Research Laboratory of Metabolic & Developmental Sciences, Department of Bioinformatics and Biostatistics, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China

***Corresponding Author**

Hai-Feng Chen (Full Professor)

State Key Laboratory of Microbial metabolism, Joint International Research Laboratory of Metabolic & Developmental Sciences, Department of Bioinformatics and Biostatistics, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China

Tel: 86-21-34204073; **Fax**: 86-21-34204073; **Email**: haifengchen@sjtu.edu.cn ; weitinging@sjtu.edu.cn

Notes

The authors declare that there is no conflict of interest.

Author Contributions

[#]These authors contributed equally to this work.

ABSTRACT

Proteins are essential macromolecules that play crucial roles in nearly every type of biological function. Most of the protein functions are determined by their position topologies, indicating that new functional proteins might be generated by designing their tertiary structures initially. Over the past two decades, numerous energy-based and machine learning algorithms have been proposed for *de novo* protein design. However, optimizing these models to achieve a balanced performance among three critical aspects, namely sample fidelity, diversity, and generation speed, remains a challenging task on account of the constrained residue connections and hydrophobic interactions. Here we used the protein language model (pLM) as supervision to develop two complementary deep learning algorithms, Generative Protein Design by Language Model-Inpainting (GPDL-Inpainting) and GPDL-Hallucination, that can sufficiently represent and generate the protein structures and sequences. We compared the models with the state-of-the-art protein design models (RFdiffusion, RF Inpainting, and RF Halluciantion) and demonstrated that our methods can generate more designable structures with more diversity and higher computation efficiency. We also explored the possibility of the combination of hallucination and inpainting which might further improve the model efficiency and diversity. The source code of GPDL is available at https://github.com/sirius777coder/GPDL.

Key Words: Protein backbone design; Protein language model; GPDL-Inpainting; GPDL-Hallucination

Introduction

Protein design has experienced remarkable progress in recent years, thanks to significant advancements in deep learning algorithms. These advancements have not only facilitated accurate prediction of protein structures[1–4] but also enabled the design of *de novo* proteins with specific functions[5–7]. Furthermore, cutting-edge methods can now be employed to design binders[8], homo-oligomers[9], active pockets[10], and antibodies[11]. Specifically, numerous protein backbones can be generated exhibiting the desired biochemical function, followed by the identification of protein sequences that can fold into these backbones[12–16]. These designs are then validated through computational methods such as self-consistency metrics[17] to check the consistency of the AI predicted structures and the designed structures. The filtered designs will be expressed to test their functions or structures in the wet lab experiment.

Lots of methods in the past two decades rely on minimizing the predefined energy functions such as Rosetta[18–20]. These methods have been successfully utilized to create a wide range of novel proteins, including binder proteins[21], transmembrane pores[22], and target proteins for diseases[23]. However, the energy-based modeling approaches face two critical challenges. Firstly, due to computational constraints and limited biochemical knowledge, the energy function can model only two-body interactions at most, resulting in inaccurate protein energy landscape modeling[24]. However, many functional regions in proteins involve complex interactions, such as hydrogen bond networks and catalytic triads, which cannot be adequately described by a two-body energy term. Secondly, the optimizing methods, such as Markov Chain Monte Carlo (MCMC), require numerous trajectories and thousands of iterations to converge in the rugged landscape of protein design. This intensive computational demand significantly challenges efficient design process. As a result, new approaches are needed to address these limitations and further advance protein design.

Since AlexNet achieved remarkable performance in the ImageNet competition in 2012[25], deep learning has undergone rapid development. The powerful representation

capability makes deep learning network highly suitable for protein design, which is an NP-hard problem involving combinatorial optimization of sequences and structures [20,26]. As for biological science, there are two different paradigms to train DL models. The first is to train the model from scratch, usually for some straightforward tasks and small models. The other way aims to fine-tune a general and pre-trained model for specific downstream tasks. Recent research has shown that although there is no guarantee that a pre-trained model improves model performance, the convergence time for fine-tuned model is usually shorter than that for training from scratch[27].

Protein backbone generation refers to the *de novo* design of protein tertiary structure that adheres to physicochemical principles. The generated backbones should also possess high structural "designability"[28], which means that amino acid sequences can fold into the designed backbone structure in their native states with relatively low energy. In comparison to fixed-backbone sequence design tasks generating sequences for the target topology structure, backbone generation is more intricate. The designed results should not only account for the invariance of Euclidean transformations, but also satisfy various complex physicochemical properties, such as peptide bond constraints.

Here we focus on the motif-scaffolding problem[6]. The main idea is that protein functions, such as binding and catalysis, are mediated by the motifs, consisting of several (usually a small number of) residues held together by the other protein backbone. Therefore, the scaffolding motif is a crucial and common problem in protein engineering. For this problem, the motif region can be identified by quantum computation or docking. In this study, we design a scaffold protein that can hold this motif with minimum total energy[29].

There are primarily three distinct deep learning methods in protein structure generation: inpainting, hallucination, and diffusion. Currently published RF inpainting method was initially fine-tuned from the RoseTTAFold[2,36] by the joint sequence and structure prediction tasks. The probability formulation is as follows, s and x represent the structures and sequences of proteins, and $\tilde{s}$ and $\tilde{x}$ denote the sequences and structures

of active sites. We use the neural network parameters θ to get the joint distribution $p_\theta(s, x)$.

$$s, x \sim p_\theta(s, x | \tilde{s}, \tilde{x}) \tag{1}$$

As RoseTTAFold originally trained by the Protein Data Bank database with multiple sequence alignment which highly relies on the evolutionary information, fine-tuning it will worsen the model performance especially on the sample diversity.

Hallucination method also relies on a well-trained protein structure prediction network. After initializing the protein sequence randomly and defining the loss functions, Monte Carlo simulated annealing is employed to optimize in the sequence space. The loss functions typically include: 1. Overall confidence of the protein structure, and 2. The loss function for active sites, which measures the deviation between the generated protein structure and the target protein structure. By maximizing the confidence and minimizing the topology difference in active sites, highly confident hallucinated structures that closely resemble the target structure can be obtained. Over the past two years, numerous proteins, including monomers, homo-oligomers, and complex pockets have been generated using hallucination-based methods, as these methods do not require retraining the network and allow for customization of the loss function.

However, this approach currently has two critical drawbacks: 1. The optimization is along the energy landscape of sequences rather than structures, which might lead to confident but non-expressed adversarial sequences[9]. 2. The loss function convergence typically involves a combination of simulated annealing and gradient descent, which cannot be accelerated using mainstream accelerators such as GPU. As a result, the convergence speed is notably slow, preventing high-throughput hallucination generation under limited computational resources.

The RFdiffusion probability generative model has attracted much attention in the past three years and it has already revolutionized the AI generated image field such as Midjourney and Stable Diffusion. It assumes the original data distribution is $x^0 \sim$

$P_\theta(x)$ and then gradually add Gaussian noise to make this data distribution to a N-dim standard Gaussian distribution $x^T \sim \mathbb{N}(x^T; \mathbf{0}, I_n)$. Here is the forward process to add the noise, $\epsilon^t \sim \mathbb{N}(\mathbf{0}, I_n)$ is independent with each other and β^t is a set of hyperparameters.

$$x^{t+1} = \sqrt{1-\beta_t}x^t + \sqrt{\beta_t}\epsilon^t \ , t \in [0,1 \dots T] \tag{2}$$

The reverse sample process was conducted by a fixed Markov process. This approach has many advantages: 1.As the model directly samples backbones from Gaussian noise, sampled proteins have high diversity. 2. The model can optimize the random noise along the structure and sequence landscape[37,38]. 3. Diffusion model can receive different types of information such as torsion angle, distance map and residue gas, which makes the generation process controllable.

However, its shortcomings are similar to hallucination. Generating a protein requires a substantial denoising process, and due to the complexity of protein conformations, diffusion models heavily rely on fine-tuning with pre-trained protein structure prediction models. It is challenging to have a lightweight model for inference purposes.

Protein language models are large pre-trained models utilized for sequence representation[39–41] and can be fine-tuned for various downstream deterministic tasks, including protein structure prediction[3,4], *de novo* sequence generation[42–44], protein function annotation[45–48], and protein-protein interactions[49]. In this study, we employed the protein language models ESM2 and ESMFold[3] to construct Generative Protein Design by Language Model (GPDL), including two distinct protein structure design models, GPDL-Inpainting and GPDL-Hallucination.

The initialization of GPDL-Inpainting is performed using the ESMFold parameters, and the model is further fine-tuned through the missing information recovery task. Specifically, the model is trained using masked sequences and masked backbone structures as inputs, with the objective of predicting the complete sequences and structures. On the other hand, GPDL-Hallucination is based on the concept of

constrained hallucination, which receives a random sequence input and predicts a noised structure. We subsequently used MCMC simulated annealing to iteratively optimize this sequence until convergence into a high-confident prediction.

There have been some works to use protein language model in backbone generation[50] and fixed backbone design[51,52]. However, in this study we focused on the motif-scaffolding problem and used our models in a series of backbone generation problems systematically to evaluate their performance. Ultimately, we underscored that our approach introduces a novel paradigm, GPDL, to address the motif-scaffold generation challenge. For cases of relative simplicity, the inpainting technique enables rapid attainment of favorable outcomes. Whereas for more intricate scenarios, such as scaffolding motifs with a few residues or in loop regions, the hallucination, coupled with MCMC optimization, can be employed to yield ideal backbone structures.

Results

Overview

In this paper, we leveraged the potential of protein language models[3] to create two distinct yet complementary and lightweight protein backbone generation tools as Fig. 1. GPDL-Inpainting on the left used motif sequence and distance map as input and generated the whole sequences and structures in a few circles (We set T = 1 in all the experiments). GPDL-Hallucination positioned on the right gets random sequence and iteratively refined this sequence to obtain a lower loss value.

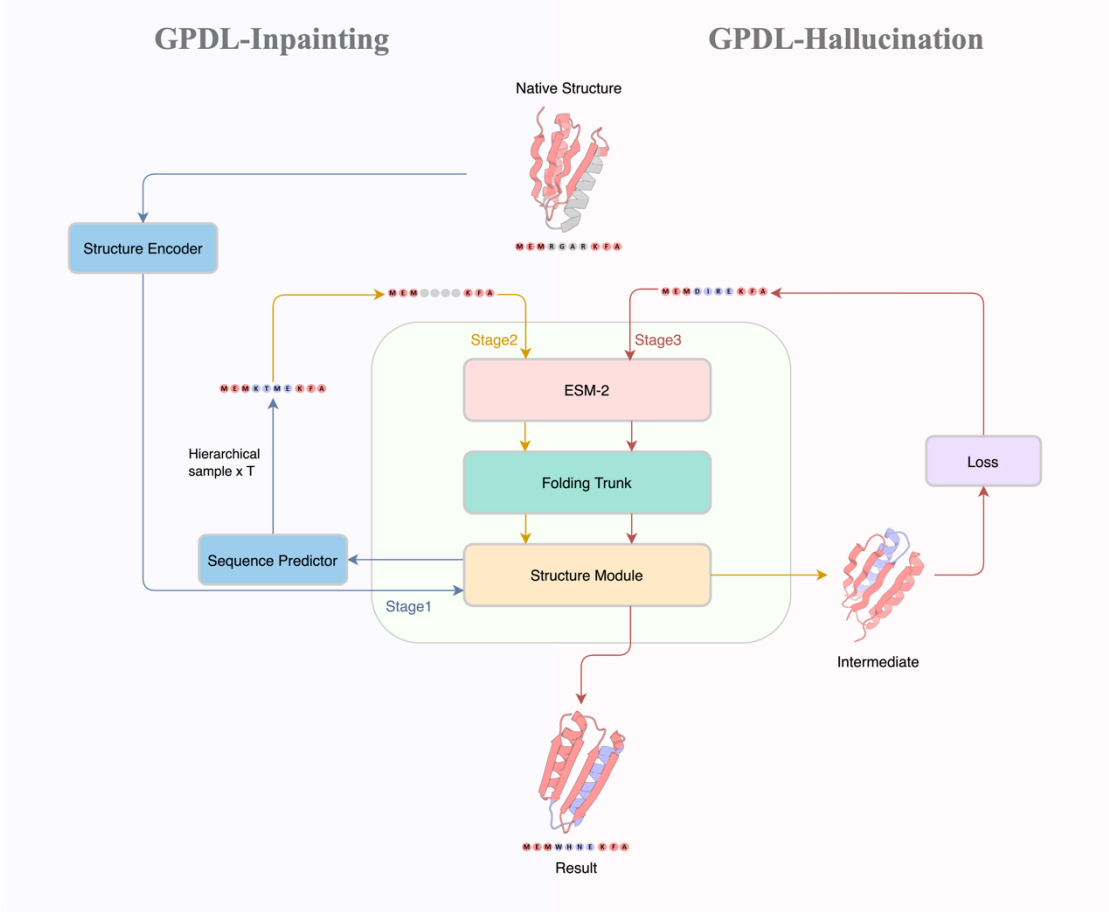

Figure 1. The framework of the model. The left (purple) represents the GPDL-Inpainting model and the right (pink) represents the GPDL-Hallucination model. The input structure in the top has the motif region (pink) and scaffold region (gray). The output in the bottom is a designed structure with the conserved motif (pink) and a new designed scaffold (blue). The PDB ID in the plot is 2KL8.

GPDL-Inpainting

Model architecture

In this study, we present a novel and computationally efficient framework derived from the ESMFold architecture, which comprises three interconnected components. The central element was to utilize the ESM2 language model to extract implicit evolutionary information from the input sequence. The resulting single sequence representation was then combined with a pair representation within the folding trunk module. Subsequently, the structure module was employed to generate the all-atom protein structure.

In order to fine-tune it for the inpainting task, we extended the original architecture with linear probing layers. These layers enabled the retrieval of input motif information and output the generated structures and sequences. Model architecture is illustrated in Figure 1 and the training details in supplementary (Fig. S1A-D).

GPDL-Hallucination

Overall framework

A critical aspect of the scaffolding problem was the preservation of functional sites or motifs, which are essential for maintaining the biological activity of protein. Achieving this objective required careful consideration of the overall structure of generated backbone, as it determined the proper placement of the motifs within the entire structure. In this regard, the constrained hallucination method had emerged as a promising approach to designing diverse protein backbones that retain their functional potential[6,58].

This constrained hallucination involved iteratively introducing mutations to an initial sequence and subsequently updating the sequence based on pre-defined criteria(Fig. S1E). Through this iterative process, a final backbone structure was generated, which exhibits optimal function potential(Fig. S1F). Generative protein design by hallucination (GPDL-Hallucination) utilized the ESMFold as a structure prediction module to generate multiple backbones based on hallucination. The protein structure prediction method ESMFold, with a large language model to extract the sequence information, avoid the dependence on searching for multiple sequence alignments (MSA) and templates. This feature made it suitable for *de novo* protein design. Moreover, its high speed and accuracy facilitated the structures optimization.

We also examined combinations involving two distinct models. we found that substituting the random sequence in hallucination with the energetically favorable sequence derived from inpainting as a starting point could potentially lead to a faster MCMC convergence, and could provide ideal results for tasks in which the inpainting

failed to complete.

Benchmark set

To rigorously evaluate our methods *in silico*, we used the benchmark set constructed by RFdiffusion[5]. We employed both GPDL-Inpainting and GPDL-Hallucination to design a variety of functional proteins, including immunogen candidates, receptor traps, metal-coordinating proteins, and other protein-binding proteins(Fig. S2, Tables S1-S2).

Besides the traditional metrics such as self-consistency, pLDDT, and RMSD, we compared the computation efficiency and sample diversity which demonstrate our methods with better performance.

Performance of fidelity

To evaluate the results, we used the benchmark to compare our method's performance with mainstream methods like inpainting and hallucination. The success rate of our method was calculated and compared to that of other methods (Fig. S3 and Tables S1-S2).

We used two methods to scaffold the EF-hand motif (Figure 2A-C) and the calcium-binding site with the secondary structure of helix-loop-helix (Figure 2D-E). In this case, both GPDL-Hallucination and GPDL-Inpainting generated backbones that accurately recapitulated the motif with a motif RMSD less than 1Å. Moreover, ESMfold prediction results demonstrated a close recapitulation of the backbones, with an overall RMSD of less than 2 Å and pAE of less than 5Å in both methods. These results illustrate the optimal reliability and functional potential of the designed structures.

In comparison with other methods, our hallucination method exhibited high success rate in scaffolding the EF-hand motif. In every roundtrip of GPDL-Hallucination, we obtained in silico successful results, as evaluated by the benchmark criteria mentioned above. This outstanding success rate of GPDL-Hallucination highlights its effectiveness in generating functionally relevant protein structures with high precision and reliability.

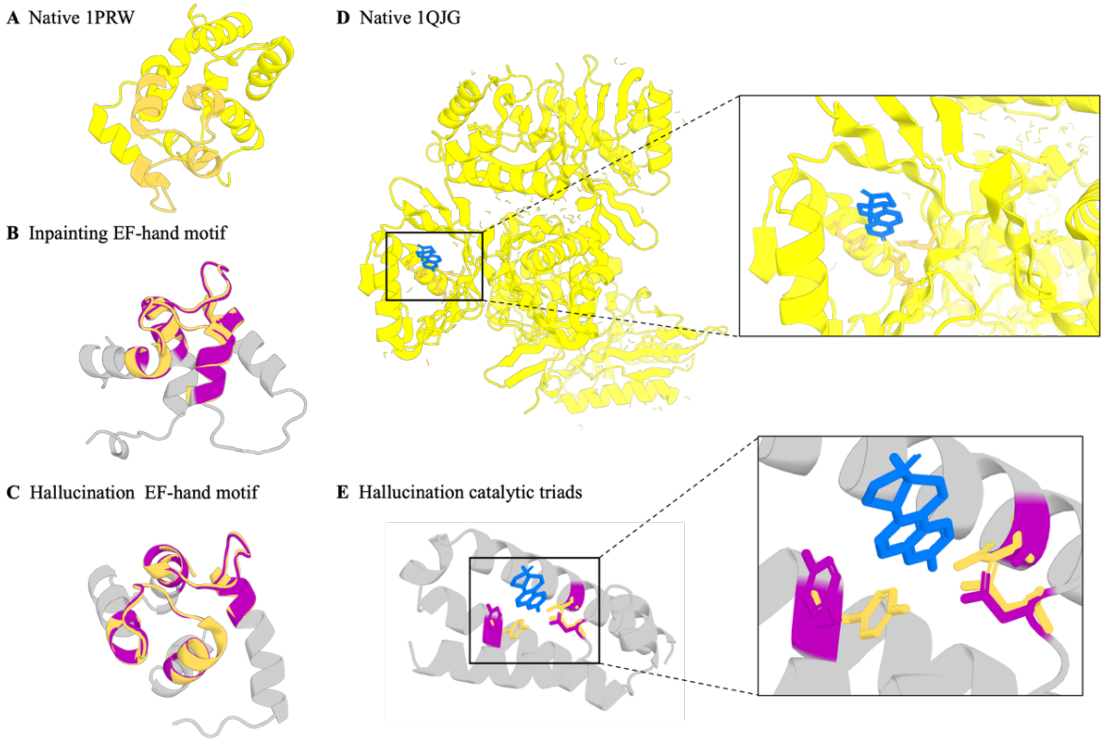

Figure 2. Design of metal-coordinating protein and enzymes. (A-C) 1PRW metal binding functional sites. (A) Native protein PDB ID 1PRW and the EF-hand motif chain A residues 16 to 35 and 52 to 71. (B) Inpainting EF-hand motif. Accuracy : motif RMSD = 0.398Å (with native motif), prediction confidence: pAE = 4.42Å, self-consistency RMSD = 1.78Å. (C) Hallucination EF-hand motif. Accuracy : motif RMSD = 0.33Å (with native motif), prediction confidence: pAE = 2.76Å. self-consistency RMSD = 0.62Å. Colors: native protein scaffold, yellow; native functional motif, yellow-orange; inpainting scaffold, gray; inpainting motif, purple. (D and E) 1QJG catalytic triads scaffolding.(D) Native protein PDB ID 1QJG and the motif A residues 16 to 35 and 52 to 71. (E) Hallucination catalytic triads. Accuracy : motif RMSD = 0.93Å (with native motif), prediction confidence: pAE = 2.66Å, self-consistency RMSD = 0.62Å.Colors: native protein scaffold, yellow; native functional motif, yellow-orange; inpainting scaffold, gray; inpainting motif, purple.

Additionally, we tested two methods on designing enzymes, focusing on the active site of KSI, which is involved in steroid hormone biosynthesis. The functional motif is a catalytic triad composed of an aspartic acid, an asparagine, and a tyrosine. This presented challenges due to the higher degree of conformational flexibility and less restriction in the extremely low number of amino acids in the catalytic active site. Notably, the scaffold created by hallucination accurately recapitulated the catalytic triad with sub-angstrom accuracy(Fig. 2E), maintaining the orientation of the side-chains while RF design cannot scaffold it[5,6].

Furthermore, the recapitulation of the backbone by ESMFold, with overall RMSD of 0.62Å and pAE of 2.66 Å confirms the functional potential and structural reliability of the design. It's worth mentioning that the generated backbone showed a different fold with the native reference, with the aspartic acid and tyrosine in α-helix, while they were in α-helix and β-sheet respectively in the native reference. The loop holding asparagine was maintained, and as the loop possesses high flexibility, this preservation contributes to maintaining the catalytic potential of the designed protein. Moreover, the smaller size of the designed backbone compared to the reference indicates the potential of our method to search for smaller backbones for functional proteins. This aspect holds significant application value in medical research and healthcare.

Performance of diversity

We generated 200 proteins by three different methods: GPDL-Inpainting, GPDL-Hallucination and RFdiffusion, respectively. To assess the sample diversity, we used two cases with different motif secondary structures. We analyzed the generation results by the dihedral distributional (Fig.3A), bond angle and bond length (Fig. S4) and find ϕ/ψ distribution agree with the standard Ramachandran plots especially in the beta-strand and alpha-helical regions.

Subsequently, we quantify the output diversity by multi-dimensional scaling(MDS) of pairwise backbone TM-score between all the designed results which has been used in the protein oligomers hallucination problems[9](Fig. 3B). TM-score (template modeling score)[59] uses heuristic progressive alignment to find the optimal alignment between two structures, and compared to RMSD, it better reflects the global conformational similarity. We integrated MDS plot with Gaussian mixture model and convex hull algorithm to explore sample diversity in different methods and demonstrate our method shows superior diversity compared with RFdiffusion.

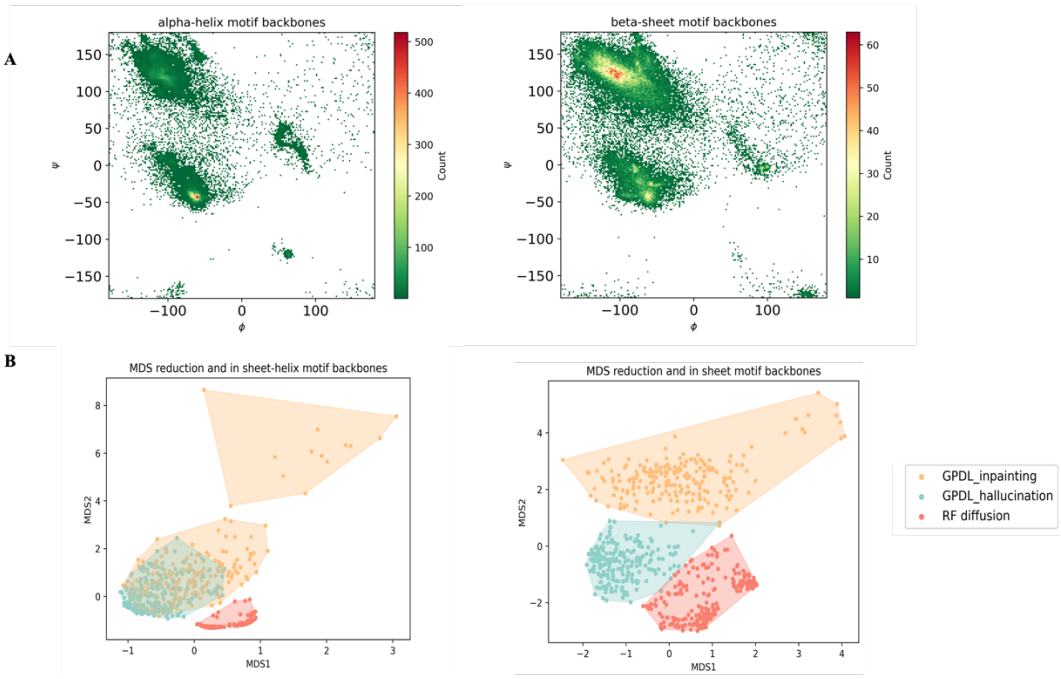

Figure 3. Backbone analysis in different motif topologies. (A) Ramachandran plot of different backbones, left column beta-sheet motif backbones (PDB ID 5IUS), right column helix-sheet mixture motif (PDB ID 2KL8). 200 inpainting and 200 hallucination backbones were mixed for each distribution plot and each backbone with same scaffold residues as RFdiffusion. (B) MDS reduction plot of inpainting, hallucination and RFdiffusion.

Furthermore, our two complementary methods show some preferences in different backbone generation tasks. In the alpha-helical motif generation problem, inpainting model could cover the hallucination region while in the other these two methods only showed tiny overlaps. In both cases, RFdiffusion showed less clustering area compared with GPDL.

We performed hierarchical clustering on the generated backbones to concretely assess sample diversity. As depicted in Fig. S5, we employed pairwise TM-score calculations utilizing a subset of samples extracted from PDB entry 2KL8 and 5IUS. The resulting clusters of 2KL8 by GPDL-Inpainting and 5IUS by GPDL-Hallucination revealed two distinct cluster centers, namely helical-loop and strand-loop conformations. These findings show that these backbones does not exhibit a bias towards helical backbone generation, although helical secondary structures represent relatively stable motifs

within the protein data bank. However, the cluster of 2KL8 by GPDL-Hallucination show that most of the result are in the secondary structure of helical-loop. The helical bias was also observed in some intricate scenarios, such as scaffolding motifs in loop regions. As this bias may lead to a destruction in recapitulation of the motif, the causes and way to prevent is still waiting for further research.

In the previous research, hallucination and diffusion techniques often exhibited greater diversity when contrasted with inpainting. Nonetheless, our GPDL-Inpainting approach demonstrates superior sample diversity, which can be attributed to the protein language model that has explored lots of different sequence patterns in the training processes. We hypothesized that as the proportion of the masked sequence region increases, the motif embeddings exhibit a greater degree of differentiation.

Sampling quality with different sequence lengths

The protein scaffold minimization presents an intriguing challenge in protein engineering – can we achieve the shortest possible protein backbones for the target motif part? However, while shorter protein backbones exhibit lower energies and more stable conformations, they might not preserve the targeted function. In this section, we investigated the quality of generated backbones with distinct secondary structures across various length scales, aiming to identify the optimal length interval with relatively high prediction confidence, lower motif geometry difference and shorter scaffold length.

Utilizing GPDL-Inpainting, we sampled 200 designs filtered with RMSD < 5.0 Å. Subsequently, we categorized the proteins into 7 groups based on their lengths and plotted the kernel density for their pLDDT and motif RMSD. Remarkably, we observed that the motif RMSD remains relatively unaffected by the length interval, whereas shorter lengths generally yield more confident backbones. To define an ideal range of protein lengths that encompasses regions of high prediction confidence (exceeding 10%) and low motif structure deviation (RMSD < 1.0 Å), while maintaining a short scaffold length, we concentrated on three distinct motif topologies, each comprising approximately 10 to 20 motif residues. Our analysis unveiled that, for the helix-loop motif (Fig. 4A), the total length spanning from 50 amino acids to 75 amino acids satisfies the specified criteria. Furthermore, we identified that a total length ranging from 75 amino acids to 100 amino acids represents a favorable interval that optimally strikes a balance between confidence and accuracy for the beta-strand motif (Fig. 4B) and the helical-sheet mixture motif (Fig. 4C). This outcome demonstrates that the helical motif typically requires a shorter scaffold length to preserve its activity, while the beta-strand topology tends to demand a longer scaffold length to achieve the same outcome.

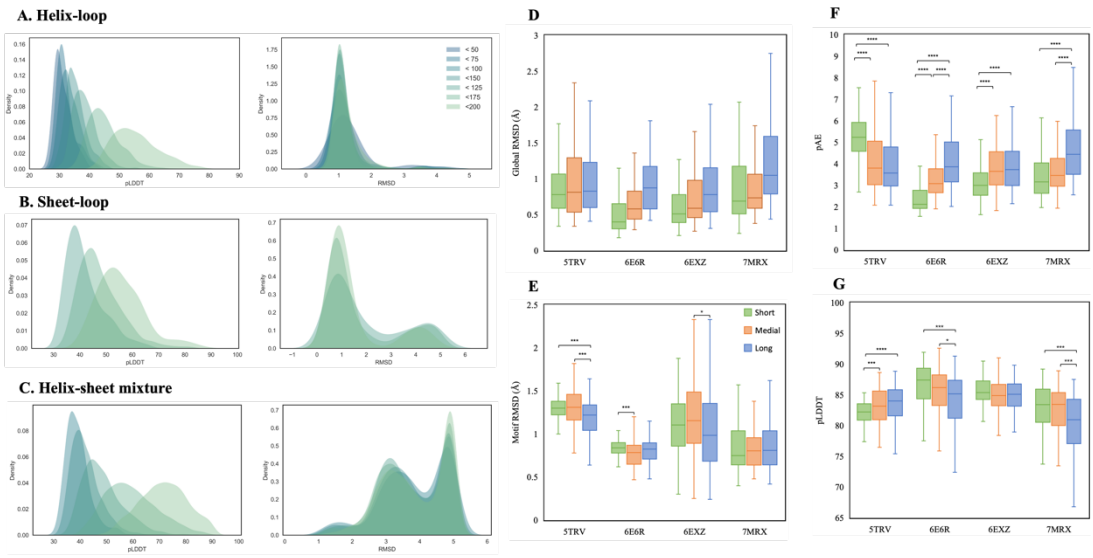

Figure 4. Backbone generation distribution in different motif topologies and length groups. (A) helix-loop motif, PDB ID 6E6R. (B) sheet-loop motif, PDB ID 4ZYP. (C) helix-loop-sheet mixture motif, PDB ID 5TRV. Left column : generated pLDDT, right column : motif RMSD with native backbones. (D) Global RMSD, motif RMSD, pAE and pLDDT of designs by hallucination, PDB ID 6E6R, 6EXZ, 7MRX (helix-loop motif) and 5TRV (helix-loop-sheet mixture motif). *, P-value < 0.05; **, P-value < 0.01; ***, P-value < 0.001; ****, P-value < 0.0001.

The performance described above is also applicable to the hallucination method (Fig. 4D-G). In this context, we selected four proteins with distinct secondary structures and employed the hallucination to generate backbones of varying lengths. Subsequently, we conducted a comprehensive analysis of these backbones, considering four metrics: global RMSD, self-consistency RMSD, pAE and pLDDT. The results of the hallucination indicated that the length of the backbones did not exert a significant influence on the accurate reconstruction of motifs and the folding patterns of the redesigned sequences. Notably, the average pLDDT scores for all generated backbones exceeded 80, signifying high overall quality. However, it was observed that shorter backbones tended to exhibit greater confidence in their predictions.

Remarkably, the performance of GPDL-Hallucination in the case of the PDB 5TRV task displayed distinct characteristics when compared to other tasks. We attributed this divergence to the unique structure of the motif in 5TRV, characterized by a bending

motif comprising a helix and a sheet connected by a loop. Consequently, such a structure necessitates a longer scaffold for adequate support during the hallucination process.

Performance comparison

We evaluated our designs generated by GPDL-Inpainting and GPDL-Hallucination using the benchmark criteria and compared them with designs produced by RF design, and scaffolds encoded by random sequences. This comprehensive evaluation allowed us to assess the performance of our methods in comparison with others.

Compared to backbones generated by random sequences, those produced by GPDL-Inpainting and GPDL-Hallucination demonstrated a remarkable ability to accurately recapitulate the conformation of the motif during the design process. Furthermore, our methods exhibited consistent and stable performance across all scaffolding tasks in our dataset, which encompassed a variety of secondary structures. In over half of the designs, the motif was recapitulated with sub-angstrom accuracy, indicating the high precision achieved by our methods (Fig. 5A).

Additionally, the designs generated by GPDL-Inpainting and GPDL-Hallucination exhibited relatively low global-RMSD values(Fig. 5B), signifying the high self-consistency of the backbones. Moreover, these backbones could be effectively encoded by the redesigned sequences. The structure prediction of our designs also yielded high pLDDT scores(Fig. 5C) and low pAE scores(Fig. 5D), which are measures of the confidence in the prediction results. The combination of high confidence and self-consistency in the backbones further supports the high reliability of the backbones provided by our methods.

In this section, we followed the definition for successfully designed proteins before. Our inpainting model achieved a success rate of 20 out of 25 attempts, while the hallucination method achieves success in 23 out of 25 attempts, same to the success rate of RFdiffusion.

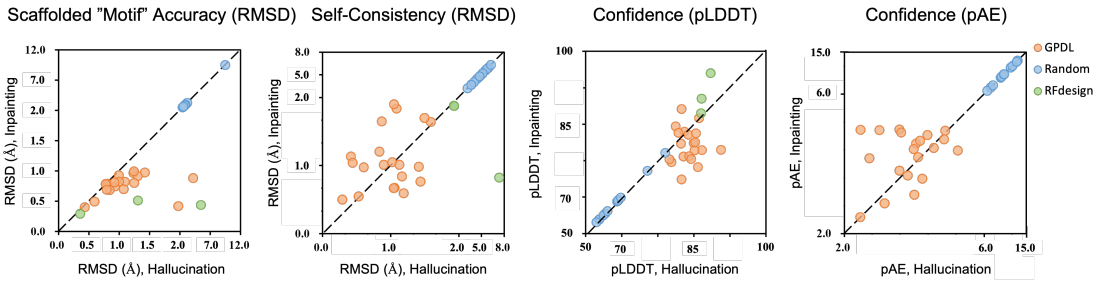

Figure 4. Benchmark plots of 25 different cases. (A) motif RMSD, designed backbone motif with native motif. (B) global RMSD, designed backbones with ESMFold prediction backbones. (C) pLDDT predicted by ESMFold. (D) pAE predicted by ESMFold.

It is worth emphasizing that the inpainting method demonstrates notably efficient results within a remarkably short time, as demonstrated in Table 1. Furthermore, we compared our inpainting methods with RFdiffusion in the same table. we executed the tests on a single Tesla V100 GPU, generating a diverse set of conditioned proteins with varying lengths. The last two columns in the table represent the run time of the different models for generating a single sample. Each method produced 100 conformations for different protein systems and the average time was recorded. The results reveal that our methods achieved an impressive nearly 40-fold speedup for the same protein system with a length of 100 amino acids compared to RFdiffusion. We ignored the runtime of hallucination as each result calls for hours of MCMC iterations.

Table 1. Computation efficiency in different length

PDB ID	Sequence length	GPDL-Inpainting(s)	RFdiffusion(s)
1YCR	29	2.18	34.2
2KL8	79	2.59	74.4
7MRX	98	2.60	70.2
7MRX	148	3.57	39.0

7MRX	265	11.32	135.6

Model combination

While GPDL-Inpainting demonstrates superior computational efficiency compared to other tools, we have observed its diminished accuracy in intricate motif topologies.(Fig. S2). As part of our methodology, we employed the inpainting-generated sequence as the starting point for the GPDL-Hallucination program, and constructed GPDL. In this section, we tried to generate the scaffold for RSV G-protein 2D10 site (PDB ID: 5WN9), in which our inpainting model failed to generate a designable backbone structure recapitulating the functional motif. After hundreds of simulated annealing steps in the GPDL-Hallucination, the generated structures become more compact and highly confident with preserved motif structures. Notably, the motif RMSD saw a substantial decreased from 5.0Å to 0.66Å (Fig. S6A). Furthermore, GPDL exhibited a remarkable capability in expediting the convergence of MCMC during hallucination and generating backbones efficiently. When the intermediate of GPDL first met the predefined criteria (at step 1257, with motif RMSD at 0.87Å and pLDDT at 80.0), the original GPDL-Hallucination trajectory had not yet attained the desired state (with RMSD at 0.66 Å and pLDDT 70.1), primarily due to insufficient optimization in the loop region (Fig. S6B). As for the GPDL-Hallucination, while it can offer promising scaffolds, the success rate for this case was observed to be lower than 20% even with optimization steps up to 1500 (Fig. S3). This observation suggests that the combination of two methods provides an ideal way to design backbone structures efficiently.

Discussion

Innovation and future improvement

In this study, we leveraged the advantages of language models to develop two lightweight and complementary methods for protein backbone generation and motif scaffolding problem. Our results demonstrated the advantages of protein language model in protein design. The inpainting module is fine-tuned from the ESMfold and achieve high computation efficiency and diversity. The hallucination module utilized the MCMC simulated annealing and target loss function to optimize the random protein sequences which can scaffold a more complex motif part compared with inpainting. Our GPDL benefit from the highly accelerated and pre-trained ESM2 protein language model when compared with other AlphaFold-based or RoseTTAFold-based models. Previous studies have shown that the protein language model can effectively learn the deep patterns in protein sequence and structures, making them well-suited for protein structure prediction, enzyme function annotation, and protein sequence generation. In this study, we extended the application of language model to the joint generation of functional protein backbone and sequence.

Although our methods show dramatic improvement in computational efficiency and sample diversity, there are still some limitations in our method. Firstly, in this study we considered to generate protein structures with target motif structures and sequences. However, many biological functions do not have the corresponding high affinity motif pattern, which means we have to use quantum chemistry or docking to find the native active sites in the candidate protein families. Secondly, current methods cannot consider the small molecule into designed process. For inpainting model, just like AlphaFold2 and ESMFold, current structure-prediction-based algorithms cannot recognize these non-canonical inputs. For hallucination model, we tried to add the Van der Waals radius of ligands in the loss function to preserve the location of the binding molecule. We found this extra loss function indeed made the generated structures especially the side-chain atoms away from the ligand but it does not enhance the generation accuracy and

confidence. In the next stage, we will investigate how to explicitly add the function information into design process. Lastly, we will try to make the designed structures more controllable with more prior knowledge such as secondary structure.

Research prospect

Our exploration of motif scaffolding has sparked several research interest for future studies. Firstly, since the significance of backbone generation lies in obtaining smaller-sized and more stable proteins while excluding other potential functional sites to ensure a more specific function, exploring the optimal length of the sequence becomes essential. In this study, we investigated the quality of the backbones with different lengths. Our evaluation on the performance of backbones with varying sequence length on motif recapitulation and structure confidence represents an initial exploration of the problem of minimal backbone. We also recommend some detailed sequence length for a targeted motif shape and length (Fig. 4). Secondly, we observed a tendency in our experiments where the generate results exhibited a biase towards the secondary structure of helices. Understanding the cause of this bias and leveraging or avoiding it properly based on specific design task can help us obtain more suitable backbones more efficiently (Fig. 2).

Our study shows the profound impact of deep learning and protein language model on protein design. Recently there are many novel and powerful models from the pre-trained language models such as CLEAN for enzyme function annotation and CoevolveML for protein-protein interaction and coevolution. Furthermore, as the AlphaFold database and ESMFold database have already covered the entire protein universe, including the metagenomic universe, now we can use protein structure to get a pre-trained model for the downstream tasks. For instance, ESM_IF[13] uses 12 million AlphaFold distillation dataset for the fixed backbone design problem and GearNet[60] for an effective structure-based protein encoder.

In addition to being pretrained on single-state protein structures, certain methods directly learn distributional ensembles from molecular dynamics simulation trajectories,

which exhibit strong correlations with our proposed methods[7,61,62]. However, these models often show suboptimal performance when applied to structured proteins, while demonstrating relatively high sample fidelity in handling intrinsically disordered proteins (IDPs), which inherently possess greater flexibility compared to structured proteins[63,64]. All these methods show the promising aspects of deep learning, pre-trained and fine-tuning fashion for this biology problems. However, we should also be aware that many methods only show plausible results in silico with no experiment evidence. Further validation of all these computational methods in the wet lab is still necessary.

In the field of protein design, recent advancements in deep learning tools have demonstrated their complementarity with Rosetta, a widely-used physics-based approach. As many deep learning methods may exhibit lower robustness when compared to physical-based methods, several studies have successfully employed a combination of deep learning models with Rosetta to achieve more plausible protein designs. This integrative approach capitalizes on the strengths of both methodologies, leading to improved design outcomes and enhancing our understanding of protein engineering[65–67].

Conclusions

This study introduces two innovative frameworks, namely GPDL-Inpainting and GPDL-Hallucination, which facilitate the generation of entirely new scaffold proteins by incorporating activation site information in distinct scenarios. Leveraging the protein language model, our methods enhance the process of structure generation, demonstrating competitive accuracy, enhanced diversity, and heightened confidence across various motif topologies. Additionally, our approaches exhibit superior computational efficiency when compared to other MSA-based neural networks across different protein lengths.

As a result, these two methodologies offer an efficient means of sampling proteins and searching for an optimal backbone when confronted with challenging scaffolding

problems. Sequentially applying these methods can significantly elevate the design efficiency and success rate, leading to noteworthy advancements in protein engineering. Moreover, we can sequentially integrate two methods to ensure the quality of protein structures while simultaneously enhancing design speed and optimization efficiency.

Materials and Methods

GPDL-Inpainting

Model architectures

The model input consists of two parts: firstly, the language model is fed with active site sequences and alanine padding mask, while the structure module takes the input of the distance matrix of active site residues in three-dimensional space. The distance matrix information includes the distances between each pair of amino acids for the 4 backbone atoms (N, CA, C, O) and 1 virtual atom, Virtual CB, totaling 25 distance measurements. The definition of the CB atom is as follows:

$$CB = -0.58273431 * a + 0.56802827 * b - 0.54067466 * c + CA \quad (3)$$

$$b = CA - N, c = C - CA, a = cross_product(b, c) \quad (4)$$

In this study, we adopt the Gaussian radial basis function to convert each distance into a 16-dimensional vector, following the same approach as employed in ProteinMPNN[12]. The L2 distance between two distinct atoms, denoted as m and n, located in separate residues i and j is defined as follows:

$$d(i_m, j_n) = \sqrt{\left(i_{m_x} - j_{n_x}\right)^2 + \left(i_{m_y} - j_{n_y}\right)^2 + \left(i_{m_z} - j_{n_z}\right)^2} \quad (5)$$

Then this distance will be transformed by 16 Gaussian radial functions (mean value from 2Å to 22Å)

$$r < d(i_m, j_n) >_k = exp^{-\left(\frac{d(i_m, j_n) - k}{\frac{22-2}{16}}\right)^2}, \forall k \in [2.0, 3.33, 4.667, \ldots, 22.0] \quad (6)$$

Primary advantage of utilizing distance matrix is to achieve rotation and translation

invariance, which provides valuable three-dimensional protein prior knowledge. Any SE3 transformation T for the input structure $\tilde{x}$ will not change the final output.

$$s, x \sim p_\theta(s, x|\tilde{s}, \tilde{x}) = p_\theta(s, x|\tilde{s}, T \circ \tilde{x}) \tag{7}$$

GPDL-Inpainting preserves most of the default configurations in ESMFold. The language model is esm2_t36_3B_UR50D version, folding trunk and structure module adopt esmfold_v1 version. We use an additive adapter to add the embedded distance matrix d_{ij} into the pair representation z_{ij} as follows. W, b represents the weight and bias in a single layer feed-forward network.

$$FFN(d_{ij}) = W_1 d_{ij} + b_1 \tag{8}$$

$$z_{ij} = z_{ij} + FFN(d_{ij}) \tag{9}$$

In the last recycle the single sequence representation s_i will be projected into the probabilities of 20 amino acid. This module sequence predictor is defined as follow.

$$FFN(s_i) = W_2 s_i + b_2 \tag{10}$$

$$softmax(s_i) = \frac{e^{s_i}}{\sum_{j=1}^{20} e^{(s_i)_j}} \tag{11}$$

$$s_i = \log softmax(FFN(s_i)) \tag{12}$$

Dataset and loss function

Inpainting model employed the CATH4.3[53] protein structure database as the dataset, ensuring that the redundancy among sequences did not exceed fourteen percent, which aligns with the ESM_IF[13] training dataset. During model training, for each protein sample, a random masking ratio between 0.3 and 0.6 was chosen, and accordingly, a portion of the sequences was masked with the alanine token, while some random chosen structural positions were masked with coordinates zero. As the sequence masking and structural masking are independent processes, we used a hybrid loss function to

combine the global protein sequence cross-entropy loss and the frame-aligned point error (FAPE) for protein coordinates. The formulas are presented as follows:

$$\mathcal{L} = \frac{1}{|B|}\sum_{i=1}^{|B|}[\beta_1 FAPE(x_i, \widetilde{x_i}) + \beta_2 CSE(s_i, \widetilde{s_i})] \tag{14}$$

FAPE loss is initially proposed by AlphaFold2[1] for protein structure prediction and demonstrate good performance in many downstream tasks such as protein diffusion models. The main idea is to align all the rigid groups and atoms between prediction results and target proteins. Here we only use one backbone rigid group and 3 atoms N, CA and C for each residue as follow. $T_i, \widetilde{T_j}$ represents the backbone frame for the predicted residue i residue and the target residue j .

$$FAPE\left(x_{i,m}, \widetilde{x_{j,n}}\right) = \sqrt{\left|T_i^{-1} \circ x_{i,m} - \widetilde{T_j}^{-1} \circ \widetilde{x_{j,n}}\right|^2} \tag{15}$$

In this study, due to computational limitations we only fine-tune the structure module which is nearly 2 million parameters. We use Pytorch[54] distributed data parallel and 4 graphics processing units (GPU) for the larger batch size. All the feature extraction from the PDB file is written by Biotite[55], Numpy[56] and Biopython[57].

GPDL-Hallucination

Loss function

For a given functional motif, GPDL-Hallucination utilizes random flanking sequences as scaffolds and initiates the design by introducing the initial sequence into a cyclic optimization process. During the optimization process, the initial sequence undergoes the structure prediction module ESMFold. Since the prediction result at this stage stems from a random sequence, the initial prediction often yields an unfavorable disordered structure. To guide the design, a loss function is defined taking the desired property of the designs into account.

In this study, two key factors were considered when determining the loss function: the maintenance of motif conformation and the reliability of the generated backbone. The

loss function is formulated as follows:

$$\text{Loss} = 100 - \text{pLDDT} + 10 \times \text{RMSD} \tag{16}$$

where the pLDDT (predicted Local Distance Difference Test) represents a measure of the confidence of the structure prediction, assessing the reliability of the designed backbone, and the Root Mean Square Deviation (RMSD) measures the root square mean deviation between the functional sites of the designed backbone and of the native structure.

By incorporating both pLDDT and RMSD in the loss function, the optimization process can effectively give protein backbones that not only retain the desired motif conformation but also exhibit higher reliability. This comprehensive approach ensures that the designed proteins are more likely to maintain their intended functionality while possessing a different structure to the native functional protein.

MCMC Optimization

The optimization procedure guided by the loss function is based on MCMC (Markov Chain Monte Carlo) simulated annealing. After the initial round of structure prediction, random mutations are introduced into the scaffold sequence, and the structure prediction is performed on the updated sequences. The acceptance of updated sequence is according to the metropolis criterion:

$$\text{n} < e^{(-\frac{L_{new}-L_{old}}{T})} \tag{17}$$

Where n is a value randomly sampled between 0 and 1, L_{old} and L_{new} are the losses calculated for the predicted results before and after the current round of mutation, and T is determined as:

$$\text{T} = (e^{\ln 0.5/500})^{step} \tag{18}$$

Where the step represents the current optimization round and the total number of optimization cycles for experiments in this study is determined as 1500. MCMC allows non-differentiable constraints to be a component of the loss function, making GPDL-

Hallucination applicated to other complex protein design problems beyond functional motif scaffolding.

As the loss function decreased, the designed backbone had optimized. After the final round of update and structure prediction, the optimized result was output. Through this interative optimization process, our model provided diverse backbones with high accuracy of the scaffolded functional motif.

Sequence design and evaluation

While both methods were effective in generating the backbones and providing corresponding sequence, ProteinMPNN[12] was used to refine the sequences corresponding to the acquired backbones, aiming to enhance the overall success rate during the subsequent evaluation process. It is important to note that during fixed backbone sequence design, all active site positions are kept fixed.

To evaluate the generated protein structures and sequences,we introduced the criteria for defining a successfully generated protein structure by three essential metrics, which includes a motif RMSD of less than 1.0Å, a pAE score of less than 5Å, and a global RMSD for self-consistency in ESMFold of less than 2.0Å. Notably, our definition slightly deviates from the original method[5], which employs a single sequence AlphaFold prediction for self-consistency. We hypothesized that current state-of-the-art single sequence prediction model ESMFold show better performance than AlphaFold2 with no MSA information. By collectively applying these evaluation metrics, we gain valuable insights into the quality and reliability of the generated protein structures, enabling a comprehensive assessment of their performance.

ACKNOWLEDGMENTS

This work was supported by the Center for HPC at Shanghai Jiao Tong University, and the National Key Research and Development Program of China (2020YFA0907700), the National Natural Science Foundation of China (21977068 and 32171242), and the

Fundamental Research Funds for the Central Universities (YG2023LC03).

References

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

2. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

3. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

4. Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *bioRxiv* (2022).

5. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* 1–3 (2023).

6. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).

7. Eguchi, R. R., Choe, C. A. & Huang, P.-S. Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *PLoS computational biology* **18**, e1010271 (2022).

8. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nature Communications* **14**, 2625 (2023).

9. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

10. An, L. *et al.* Hallucination of closed repeat proteins containing central pockets. *bioRxiv* (2022).

11. Luo, S. *et al.* Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures. in *Advances in Neural Information Processing Systems* (eds. Koyejo, S. et al.) vol. 35 9754–9767 (Curran Associates, Inc., 2022).

12. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

13. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. in *Proceedings of the 39th International Conference on Machine Learning* (eds. Chaudhuri, K. et al.) 8946–8970 (PMLR, 2022).

14. Mu, J. *et al.* De novo Protein Sequence Design Based on Deep Learning and Validation on CalB Hydrolase. *bioRxiv* (2023) doi:10.1101/2023.08.01.551444.

15. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) (Curran Associates, Inc., 2019).

16. Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A. & Kim, P. M. Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Systems* **11**, 402-411.e4 (2020).

17. Trippe, B. L. *et al.* Diffusion Probabilistic Modeling of Protein Backbones in 3D for the motif-scaffolding problem. in *The Eleventh International Conference on Learning Representations* (2023).

18. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. in *Methods in enzymology* vol. 487 545–574 (Elsevier, 2011).

19. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

20. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).

21. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

22. Xu, C. *et al.* Computational design of transmembrane pores. *Nature* **585**, 129–134 (2020).

23. Silva, D.-A. *et al.* De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019).

24. Ovchinnikov, S. & Huang, P.-S. Structure-based protein design with deep learning. *Current Opinion in Chemical Biology* **65**, 136–144 (2021).

25. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* (eds. Pereira, F., Burges, C. J., Bottou, L. & Weinberger, K. Q.) vol. 25 (Curran Associates, Inc., 2012).

26. Pierce, N. A. & Winfree, E. Protein design is NP-hard. *Protein engineering* **15**, 779–782 (2002).

27. Ahdritz, G. *et al.* OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv* (2022) doi:10.1101/2022.11.20.517210.

28. Huang, B. *et al.* A backbone-centred energy function of neural networks for protein design. *Nature* **602**, 523–528 (2022).

29. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).

30. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).

31. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. in *Advances in Neural Information Processing Systems* (eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) (Curran Associates, Inc., 2020).

32. Saharia, C. *et al.* Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. in *Advances in Neural Information Processing Systems* (eds. Koyejo, S. et al.) vol. 35 36479–36494 (Curran Associates, Inc., 2022).

33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. (2022).

34. Ingraham, J. *et al.* Illuminating protein space with a programmable generative model. *bioRxiv* 2022–12 (2022).

35. Lee, J. S., Kim, J. & Kim, P. M. Score-based generative modeling for de novo protein design. *Nature Computational Science* 1–11 (2023).

36. Baek, M. *et al.* Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv* 2023–05 (2023).

37. Anand, N. & Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019* (2022).

38. Frank, C. J. *et al.* Efficient and scalable de novo protein design using a relaxed sequence space. *bioRxiv* 2023–02 (2023).

39. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* e2016239118 (2021).

40. Rao, R. M. *et al.* MSA Transformer. in *Proceedings of the 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) 8844–8856 (PMLR, 2021).

41. Elnaggar, A. *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **44**, 7112–7127 (2021).

42. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications* **13**, 4348 (2022).

43. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* (2023).

44. Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N. & Madani, A. Progen2: exploring the boundaries of protein language models. *arXiv preprint arXiv:2206.13517* (2022).

45. Bernhofer, M. & Rost, B. TMbed: transmembrane proteins predicted through

language model embeddings. *BMC bioinformatics* **23**, 326 (2022).

46. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology* **40**, 1023–1025 (2022).

47. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).

48. Qiao, Y., Zhu, X. & Gong, H. BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics* **38**, 648–654 (2022).

49. Yang, A. *et al.* Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. *Science* **381**, eadh1720 (2023).

50. Hie, B. *et al.* A high-level programming language for generative protein design. *bioRxiv* (2022).

51. Verkuil, R. *et al.* Language models generalize beyond natural proteins. *bioRxiv* 2022–12 (2022).

52. Zheng, Z. *et al.* Structure-informed language models are protein designers. *bioRxiv* 2023–02 (2023).

53. Orengo, C. A. *et al.* CATH–a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1109 (1997).

54. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) (Curran Associates, Inc., 2019).

55. Kunzmann, P. & Hamacher, K. Biotite: a unifying open source computational biology framework in Python. *BMC bioinformatics* **19**, 1–8 (2018).

56. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

57. Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (2009).

58. Norn, C. *et al.* Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017228118 (2021).

59. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**, 702–710 (2004).

60. Zhang, Z. *et al.* Protein Representation Learning by Geometric Structure Pretraining. in *International Conference on Machine Learning* (2022).

61. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).

62. Zheng, S. *et al.* Towards Predicting Equilibrium Distributions for Molecular Systems with Deep Learning. *arXiv preprint arXiv:2306.05445* (2023).

63. Zhu, J.-J., Zhang, N.-J., Wei, T. & Chen, H.-F. Enhancing Conformational Sampling for Intrinsically Disordered and Ordered Proteins by Variational Autoencoder. *International Journal of Molecular Sciences* **24**, 6896 (2023).

64. Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nature Communications* **14**, 774 (2023).

65. Ferruz, N. *et al.* From sequence to function through structure: Deep learning for protein design. *Computational and Structural Biotechnology Journal* (2022).

66. Liu, H. & Chen, Q. Computational protein design with data-driven approaches: Recent developments and perspectives. *WIREs Computational Molecular Science* **n/a**, e1646 (2022).

67. Wang, J. Protein sequence design by deep learning. *Nature Computational Science* **2**, 416–417 (2022).