Scaffold-Lab: Critical Evaluation and Ranking of Protein Backbone Generation Methods in A Unified Framework

Zhuoqi Zheng[1]†, Bo Zhang[1]†, Bozitao Zhong[1], Kexin Liu[1], Jinyu Yu[1], Zhengxin Li[1], Junjie Zhu[1], Ting Wei[1]*, Hai-Feng Chen[1,2]*

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, 200240, China.

[2]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

De novo protein design has undergone a rapid development in recent years, particularly in fix-backbone sequence design and backbone generation. The latter stands out as more challenging yet valuable, offering the ability to design novel protein folds with fewer constraints. However, there is an absence of a comprehensive delineation of its potential for integration into practical applications in the field of protein engineering, as well as a standardized evaluation framework to accurately assess the diverse methodologies within this field. Here, we proposed Scaffold-Lab benchmark focusing on evaluating unconditional generation across metrics such as *designability*, *novelty*, *diversity*, *efficiency* and *structural properties*. We also extend our benchmark to include the motif-scaffolding problem, demonstrating the applicability of these conditional generation models. Our findings reveal that *FrameFlow* and *RFdiffusion* in unconditional generation and *GPDL* in conditional generation showcased the most outstanding performances. All data and script will be available at https://github.com/Immortals-33/Scaffold-Lab.

Key Words: protein design, protein backbone generation, evaluation, benchmark

Introduction

Proteins are fundamental to biological systems due to their diverse functions, largely attributed to the hierarchical organization that progresses from amino acid sequences to three-dimensional (3D) structures[1, 2]. Despite substantial kinds of proteins already exist in nature, they have just covered a small fraction of the whole potential protein sequence space[3]. To this extent, *computational protein design* has been developed to extend the functionality of proteins. Over the last few decades, the rapid development in protein design along with technological breakthroughs in protein structure prediction[4] has allowed biologists to dive deeper into the mystery of protein structures, breathing a new life into the field of protein science and protein engineering since the last few decades[5, 6].

In the early stages of *de novo* protein design, the task of designing protein sequences came up in the beginning, using physical-based methods which searched new protein rotamers through a conformational landscape[7]. Over the past decade, the rapid development of deep learning has greatly invigorated the field of computational structural biology, both in protein sequence design[8-15] and protein structure prediction[4, 16-21]. Although given the vast sequence space to be explored[3], it has been long acknowledged that proteins realize their biological functions directly by 3D structures[1, 22]. Based on this theory, *protein backbone generation* was carried out to directly reform the topologies of proteins[23, 24], aiming to treat the 3D structure of protein as the design objective. While accurate modeling the structural constraints in 3D cartesian space is challenging for physical-based methods, the rapid advancements in generative models in the last few years. For example, denoising diffusion models[25, 26] has been a game changer with their exceptional performance in modeling high-dimensional data distribution[27], which has already achieved incremental progress in several works[28-34].

Although protein backbone generation has seen a surge of interest in the field of protein design, several limitations have yet to be solved. First of all, most of those methods use a relatively "coarse-grained" way to either treat the all-atom protein structure as a backbone-only or C_α-only protein. This compression leads to the result that only the information of protein backbone, or namely a "scaffold" can be generated through these methods, where a sequence design method is later needed to find the optimal sequences that folding into the designed backbones. Secondly, since protein backbone generation is targeted for novel structure generation and an upstream step in protein engineering,

there is still lack of knowledge how these methods can be applied to real-world applications. Furthermore, due to the varying metrics and length scopes employed by different methods, the field lacks a set of benchmark for assessing different protein backbone generation methods, especially compared to the various established works in place for sequence design[35-38], protein fitness prediction[39] and molecular generation tasks[40].

Out of these considerations, we proposed a unified benchmark, Scaffold-Lab, to evaluate current protein backbone generation methods extensively. We selected seven representative protein backbone generation methods based on their performance and availability and carefully grouped them into different categories according to different lengths of proteins. We divided the problem sets into two different types. The first one, namely *unconditional generation*, is the task that generates complete protein scaffolds under no restrictions, with user-specified protein length as the only input parameter. Unconditional generation has been a proof-of-concept exploration of almost all protein backbone generation methods, for it implies the capability of how well those models learn the distribution of realistic protein structures, which is essential to assessing the methods' performances. Opposite to the first one, *conditional generation* creates proteins under certain user-specified restrictions. Though it has not been explored to a vast extent, it is an important task as it is the destination to fully-programmed roadmap of protein design, aiming to develop *de novo* proteins with properties specified by human beings. We selected the most widely-used *motif-scaffolding* task as a representative under this category, whose goal is to generate novel scaffolds around one or multiple motifs extracted from native protein structures, which usually plays a key role in specific biological functions. This strategy is particularly useful in applications like designing novel enzymes[41], binders[34, 42, 43] or vaccines[44] and has been widely explored and evaluated within the last few years[33, 34, 45-51]. Within this problem set, several computational metrics designed to evaluate these methods were recast into this evaluation framework followed by a ranking stage performed to assess the capability of different methods. In summary, our research undertakes a comprehensive assessment using a novel framework to elucidate the current state of progress and existing constraints within this domain, aiming to pave the way for its future development.

Material and Methods

Dataset

Table 1 listed the methods and their corresponding tasks selected to be evaluated in Scaffold-Lab benchmark test. Both *Genie*[52] and *FrameDiff*[53] uses the invariant point attention module brought by Jumper et al.[4] as the key component of their architecture. *FrameDiff* utilized diffusion modeling in Riemannian manifolds to perform protein diffusion in a SE(3) space, while *RFdiffusion*[34] integrates a diffusion models with a pretrained structure prediction model RoseTTAFold[20] to greatly improve the performance. *Chroma* incorporates a diffusion model and a graph neural network (GNN) to generate protein structure and sequence simultaneously, with a scaling technique to improve the sampling speed to a large extent. Both *TDS*[47] and *FrameFlow*[54] was developed upon *FrameDiff*'s architecture, with the former uses a particle filtering strategy to raise conditional sampling applied to motif-scaffolding task, and the latter replaces diffusion models with a flow-matching process to significantly elevate the efficiency and designability compared with previous works. Akin to the strategy in Wang et al.[46], *GPDL*[50] leverages the single structure prediction method ESMFold[16] combined with the constrained hallucination and inpainting strategies to generate highly designable and diverse scaffolds around given motifs.

Table 1. Methods Selected to be tested in Scaffold-Lab

Method	Released Date	Category†	Model Size (Parameters)	Tasks evaluated
Genie[52]	2023-01	Diffusion-based	4.1M	Unconditional
FrameDiff[53]	2023-03	Diffusion-based	17.4M	Unconditional
RFdiffusion[34]	2023-04	Diffusion-based	59.8M	Both
TDS[47]	2023-06	Diffusion-based	17.4M	Conditional
GPDL[50]	2023-10	ESM-based	3B	Conditional
FrameFlow[54]	2023-10	Flow-based	16.8M	Unconditional*
Chroma[55]	2023-11	Diffusion-based	18.4M	Both

*A motif-scaffolding version of FrameFlow has come out recently but the code of motif-scaffolding part has not yet released upon the released date of manuscript.
†We followed the category introduced by Zhang et al.[56].

Given the distinct application contexts and abilities to generate long sequences among

various methods, it is crucial to categorize them appropriately prior to evaluation. Since the only parameter needed to be set by users in unconditional generation task is the desired length of generated proteins, we further grouped these methods into three sets (short, medium and long) based on the capable designed length scope of each method. After the grouping stage, some representative lengths were chosen within the scope, through which we seek an impartial and comparative evaluation for different methods within each set. **Table 2** lists the information after grouping by lengths. The analysis throughout this benchmark would be under this length-grouping framework.

For the motif-scaffolding task, we selected the dataset brought by *RFdiffusion* in [34]. We excluded *6VW1*[57] from our evaluation as it pertains to a multi-chain motif-scaffolding task which some of the selected methods were not equipped to design. Consequently, this resulted in a dataset comprising 24 cases. More detailed referred to **Supplementary Table 1**.

Table 2. Length Grouping of Unconditional Generation

Group	Length	Method
Short	50, 100	Genie, FrameFlow, FrameDiff, Chroma, RFdiffusion
Medium	200, 300, 400, 500	FrameDiff, Chroma, RFdiffusion
Long	600, 800, 1000	Chroma, RFdiffusion

Unconditional Generation

The overall workflow of Scaffold-Lab is shown in **Figure 1**. We performed unconditional protein backbone generation after grouping different methods by lengths. For each unique length within each method, we generated 100 backbones and fed them into the refolding pipeline. With 100 sequences designed by ProteinMPNN and 10 sequences selected, we obtain a total number of 1000 refolded protein structures, which were mainly utilized to calculating designability. For the purpose of avoiding the sequence designer bias to the best degree, we only used refolded structures for the analysis on designability, while the novelty, diversity, efficiency and structural properties are calculated directly towards the protein backbones generated by tested

methods.

For the structure predictor, we used ESMFold[16] out of its outstanding performance on *de novo* proteins and its high computational efficiency. We also made a comparison between ESMFold and two other highly accurate protein structure prediction methods, AlphaFold2 and OmegaFold, using a subset of samples. The result of comparison can be found in **Supplementary Figure 1**.

Conditional Generation

For conditional generation, we chose *motif-scaffolding* as the key problem to be solved in conditional generation, where the problem is defined as below: Given a functional "motif" existing in native proteins, these methods can generate "scaffold" around or between these motifs, where the lengths of scaffolds can sometimes be variable and randomly sampled. In our selected dataset, the motifs of native proteins, the length scope of scaffolds as well as scaffolding ways have been already defined, so we set up the evaluation experiments to use these methods to run through this dataset to generate plenty of protein backbones, which would enter the refolding pipeline to be examined the fidelities of different methods.

We used GPDL-Hallucination (*GPDL-H*) to represent GPDL during the evaluation. Since *TDS* and *GPDL-H* both need to specify the total length of protein, we designed backbones with mostly the same length within each case, where the length was chosen around the median value of length scope defined in Watson et al.[34]. For *Chroma* and *RFdiffusion* we performed length-variable design for each case. We randomly generated 10 sequences using ProteinMPNN and fixed the type of amino acids inside the motif regions during sequence design stage in order to stabilize the motif regions and improve the overall success rate.

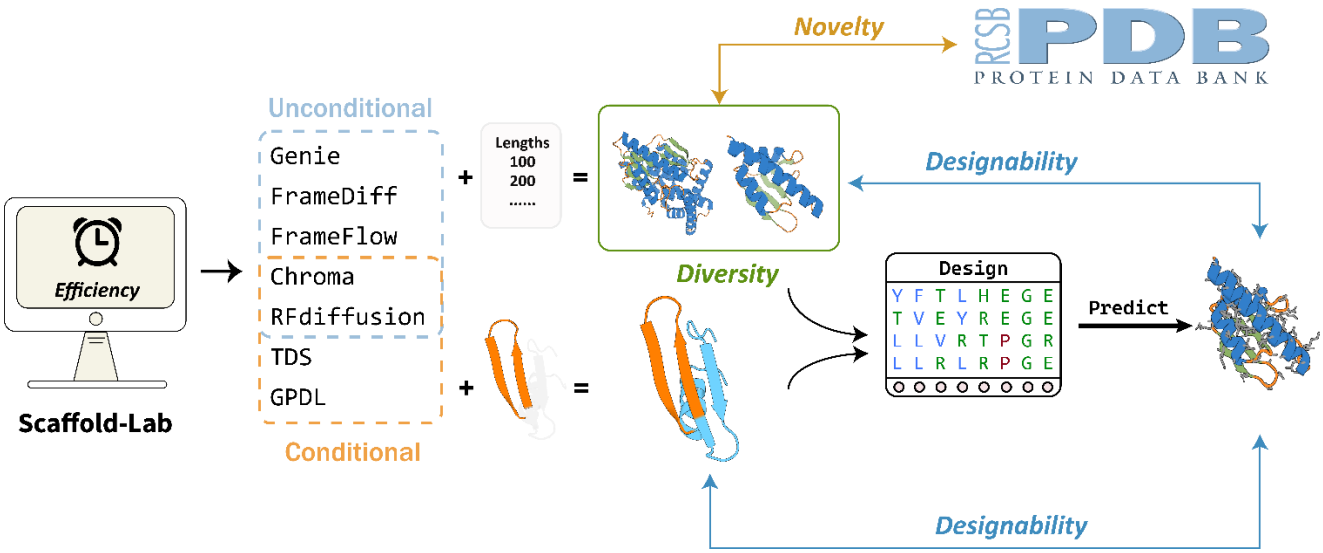

Figure 1. The overall workflow of Scaffold-Lab. Seven state-of-the-art protein backbone generation methods were evaluated by two types of tasks throughout a refolding pipeline. Four qualitative metrics were introduced to qualitatively assess the performances of methods, namely designability, novelty, diversity and efficiency.

Evaluation Metrics

In this part, we introduce the metrics used in this evaluation, encompassing a range of both quantitative and qualitative aspects. We divide these metrics into five parts, namely *designability, novelty, diversity, efficiency* and *structural properties*, where the first four qualitative metrics would be used for ranking different methods. In the following sections, we will carry out a brief introduction for each of these metrics.

Designability

Designability is the most important metric for protein backbone generation tasks as it directly measures the quality of generation, which is, the existence of any sequence capable to fold into the generated protein backbone. We used the self-consistency TM-score (*sc-TM*) metric raised by Trippe et al.[33] where a protein backbone was defined as success in designability when the TM-score[58] between refolding structures and original backbones > 0.5 as proteins with a mutual TM-score greater than 0.5 are widely considered to belong to the same fold pattern[59]. We additionally adapted two modifications:

1. We used sc-TM instead of sc-RMSD throughout our evaluation pipeline. Although many works used a threshold of sc-RMSD < 2.0 Å to define the success of backbones, we found this metric showcased commonly high values and hard for scaling during preliminary experiments, while sc-TM showed a greater robustness.

Nevertheless, we calculated the designability and success rate using sc-RMSD for a comparison and the integrity of our evaluation, which can be found in **Supplementary Figure 2**.

2. "Top N strategy". While current works considered a backbone as success when there is solely one sequence satisfy the criteria, we argue that the number of sequences that can fold into the generated structures reveal the robustness and fitness of backbones to some extent. So during our experiment, we reported the sc-TM as the mean value of N sequences with highest sc-TM among generated sequences, where the formula is defined as follows:

$$Success_{backbone} = \begin{cases} 1, \frac{1}{N}\sum_{i=1}^{N} scTM > 0.5 \mid (scTM_1 > \cdots > scTM_N > scTM_i) \\ 0, others \end{cases}$$

During our evaluation, we chose $N = 3$, which means we calculated the designability use the mean value of top 3 sequences with highest sc-TM values out of 10 selected sequences. We also reported the distribution of sc-TM and success rate by picking the best sequence as a fair comparison, which are displayed in **Figure 2B** and **Supplementary Figure 4**.

For unconditional generation, we defined the success of a protein backbone as its top 3 sc-TM > 0.5. For conditional generation (motif-scaffolding), we separated the definition for backbone success and motif success: For backbone success, we followed the criteria used in unconditional generation; for motif success, the refolded motif ought to have a motif-RMSD < 1.0 Å compared with the original input motif. The success of a scaffolded backbone was defined as backbone success plus motif success.

Novelty

Novelty is a metric that measures how novel the generated protein backbones compared to existing experimentally-solved proteins by measuring their structural similarity. We followed the novelty metric raised by Yim et al.[53] by calculating the TM-score between generated protein backbones and their most similar proteins in Protein Data Bank (PDB)[60, 61], which is referred as *pdb-TM*. If this value is low, it means that the generated backbones are novel compared to most of the existing experimentally solved native proteins, and vice versa. We use Foldseek[62] to search generated backbones against PDB to find their nearest protein structure. In order to make a fair comparison while maintaining computational efficiency, we apply the same parameter suite to all tested methods.

Apart from purely calculating pdb-TM, we hypothesis that there exists certain generated backbones with novel but undesignable structures, which is meaningless for downstream applications. Thus, we introduced a *novelty score* by adding a penalty to poorly designable backbones when calculating novelty, where it is formulated as below:

$$S_{novelty} = \begin{cases} pdbTM + \left(0.5 - w\frac{1}{\mathrm{n}}\sum_{i=1}^{n} scTM\right), \frac{1}{\mathrm{n}}\sum_{i=1}^{n} scTM < 0.5 \\ pdbTM, \frac{1}{\mathrm{n}}\sum_{i=1}^{n} scTM > 0.5 \end{cases}$$

The first parameter, n, refers to the "Top N" strategy when picking sequences generated by ProteinMPNN, while the w means the degree of penalty, where higher w brings a more stringent penalty to the less-designable backbones. We set $n = 3$ and $w = 1$ throughout our experiments. Since we define sc-TM > 0.5 as the passing threshold of designability, we leave the backbones with sc-TM > 0.5 as original pdbTM value, otherwise a penalty is put on the shortfall part between sc-TM and and the passing threshold, 0.5. We argue that after the reconsideration of novelty metric, a discrepancy between designable novel backbones and "fake novel" backbones would be exposed to a significant extent.

Diversity

Different from the two metrics mentioned above, the comparator of diversity is a group of protein backbones rather than a single protein backbone. Diversity was first designed to measure internal discrepancy of generated samples, partially reflecting the determinacy and the generalizability of models. Under the scenario of protein backbone generation, a method with greater diversity implies more diverse structural topologies can be selected as candidates for subsequent sequence design stage. This is advantageous for enhancing both the heterogeneity of potential functions of proteins and the success rate for synthesizing them during wet-lab experiments.

Since diversity is a referent metric of internal differences of generated examples, clustering the generated protein backbones is a natural way to assess the similarities between these protein backbones. We grouped the protein backbones by different methods and protein lengths and subsequently used Foldseek-Cluster[63] to cluster each group of protein backbones. We use a clustering threshold of TM-Score as 0.5 based on the aforementioned consensus[59]. For the definition of diversity, we followed Yim et al.[53] as the number of unique clusters divided by the total number of protein

backbones each group:

$$Diversity = \frac{N_{unique\ clusters}}{N_{total\ number\ of\ proteins}}$$

Under this definition, a higher count of clusters reflects more heterogeneous topologies appeared within a subset of sampled protein backbones, thus a higher diversity.

Efficiency

Denoising diffusion models have ushered a new wave in the field of protein backbone generation as a representative of generative models. Despite their outstanding capabilities, many of them suffer from the iterative time-consuming denoising steps, which makes them suffer from the expensive consumption of both time and computational resources. Since efficiency is particularly crucial for the high-throughput scenario of protein engineering, we also compared the runtime of different protein backbone generation methods by generating 100 backbones per group and calculating the average time taken for each backbone. For the fairness of comparison, we maintain the hardware resources to be the same when performing experiments with different methods by conducting all experiments on a single CPU.

Structural Properties

Proteins have various forms of internal structures organized in a hierarchical way. A protein with desired folded state usually has many unique covalent and non-covalent interactions which is a key theme in stabilizing the protein and leading it to an energy-favored state[2]. As protein backbone generation focus on the three-dimensional modeling of proteins, it is crucial to check the physical plausibility and quality from the structural perspective. In this part, we analyze the structural properties of protein from two different aspects, *secondary structure elements* and *radius of gyration*, respectively.

The secondary structure element (SSE) is a key feature of proteins since it directly represents the type of topologies and structural organization of a protein backbone. Since SSE is mainly determined by the backbone of protein and previous research has already shown that sequence design methods might bring internal bias for SSE preferences[38], we directly compute SSE for the original generated protein backbones. We calculate SSE using DSSP software[64, 65] through an API by Python package Biotite[66], count the eight types of SSEs assigned by DSSP and further group them into four different common secondary structure types, which are *α-helix, β-sheet, coil*

and *turn*, respectively. We further analyzed the proportion of different SSEs in protein backbones generated by different groups to see if there's any preference or potential rules produced by these methods.

Apart from SSEs, we also computed the radius of gyration (Rg) of each generated protein backbones. The radius of gyration of a protein can be measured in X-ray or neutron scattering or some other biological experiments, which is one of the key properties of any biological polymer molecule as it measures the size or compactness of proteins. Previous research has already shown that the value of Rg can be a key distinction between structured proteins and intrinsically disordered proteins (IDPs), whereas the fluctuation of Rg during a certain time period can reflect the flexibility of proteins to some extent, which is of great significance to the function of proteins[67]. We computed Rg directly on the protein backbones generated by tested methods as what we applied in SSE computation, and made a comparison to an experienced Rg-protein length diagram. We believe our analysis on structural properties might provide a deeper insight into the protein backbones generated by current methods.

Ranking

For the ranking of different methods, we used the widely utilized Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). TOPSIS[68] is a multi-criteria decision analysis method based on the concept that the shortest geometric distance from the positive ideal solution should be chosen as the best alternative, and vice versa. This method is usually used combined with comprehensive factor weighting methods to improve the confidence of evaluation, either subjective weighting methods like Analytic Hierarchy Process (AHP)[69] or objective methods like CRiteria Importance Through Intercriteria Correlation (CRITIC)[70]. This ranking procedure, previously used in the evaluation of protein sequence design methods[38], has been adapted for our work.

Although AHP performs well in many instances, it has been found the results based on AHP might significantly differ from the subjective weights[71]. Thus, we introduce CRITIC to objectively decide the importance of our metrics to minimize the impact of subjective factors on our ranking outcomes. After data preprocessing by setting values We With this data preprocessing procedure, we finally rank different methods based on the length groups, where the four qualitative metrics are **designability**, **novelty**, **diversity** and **efficiency**. We suggest interested readers to Supplementary Text for a

detailed version of our ranking procedure.

Results

Unconditional Generation

Designability. Figure 2 displays the results of unconditional generation task. We first reported the top 3 sc-TM distribution of different methods along different groups, which is shown in **Figure 2A**. From the result of short-protein group, nearly all the evaluated methods performed pretty well, where all the methods attained a median value above 0.7, where *RFdiffusion* even reach a strikingly high value over 0.95. Nevertheless, the performance of them dwindled as the lengths gradually increases, until neither the median sc-TM value of *RFdiffusion* nor *Chroma* reached the threshold of 0.5, though the backbone produced *RFdiffusion* has a significant proportion above the threshold of success.

To further explore the qualities of generated backbones, we calculated the success rate by both top 3 sc-TM and best sc-TM, whose results are shown in **Figure 2B**. From the trends displayed in the graph, we can observe that the success rate generally decreases as the length grows. We also noticed a watershed on the success rate: When the length is equal or less than 300, all of the tested methods show a strong performance with success rate around or more than 80%, which then rapidly dropped until less than 30% when it comes to proteins with not less than 800 residues. Additionally, the success rate between two distinct definitions indicates that the robustness of designed backbones also declines to a certain extent as the length increases.

Generally, for both the distribution of sc-TM and success rate, *RFdiffusion* performs the best among all the tested methods, while other methods are also reliable when generating short proteins. We consider all tested methods are capable of generating highly-confident proteins with residues not more than 300 residues. Nevertheless, when it comes to the situation of long proteins, it is essential to carefully check the plausibility of proteins before putting it into a downstream scene.

Out of the considerations aforementioned, we only used sc-TM as a representative metric for determining whether a protein backbone was designable or not. Nevertheless, for the integrity of our experiments, we also computed the sc-RMSD values for each method in different groups, whose results are displayed in **Supplementary Figure 3**.

As a reference, we also reported the best sc-TM distribution, which is shown in **Supplementary Figure 4**.

Novelty. We performed novelty analysis to different length groups, whose results are shown in **Figure 2C**. The pdb-TM values predominantly range between 0.5 and 0.7, with no significant differences in trends observed between various methods and length categories. Notably, *FrameFlow* and *RFdiffusion* exhibit slightly higher pdb-TM values in the short proteins group, suggesting a tendency to generate protein backbones more akin to natural proteins. Given their relatively higher success rates in corresponding categories shown in Figure 2B, we hypothesized that this trend in novelty may also reflect the fitness of model to some extent, which implies a potential trade-off between novelty and designability.

To further validate our hypothesis, we used the novelty score with penalty put on protein backbones with low designability, in order to discover the correlation between these two methods. **Figure 2E** display the differences between the distribution of pdb-TM and novelty score, where we conducted one-sided *Mann-Whitney U test*[72] across different groups distinguished by with or without the penalty term. The results show that some of the groups have significant differences with *p-values* < 0.001, including *FrameDiff*, *Chroma* in medium-protein group (6.11×10^{-8} and 5.19×10^{-4} , respectively) and *RFdiffusion, Chroma* in long-protein group (2.46×10^{-15} and 3.33×10^{-16} , respectively) . We also noticed that *RFdiffusion* and *FrameFlow* in the short protein group did not show significant differences in results with or without penalty, indicating that most proteins within those groups pass the designability threshold, which aligns with our previously proposed hypothesis. The differences between penalty groups grows along with proteins' length, indicating the appearance of protein backbones with high novelty but low designability. Our analysis suggested that the designability should be partially taken into consideration when calculating novelty, whereas different methods did not show a strong variance in novelty overall. The complete results of significant test are shown in **Supplementary Table 2**.

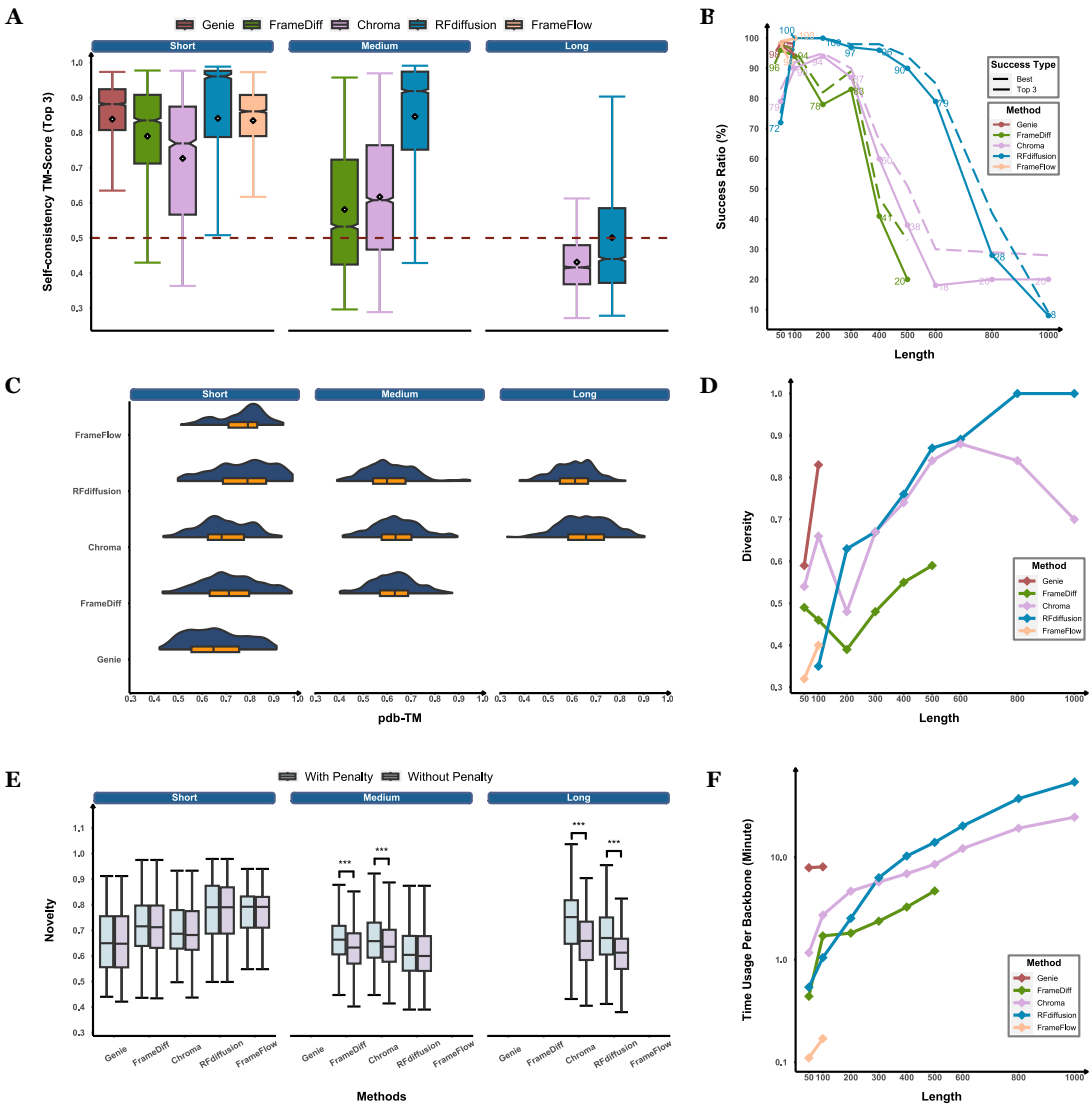

Figure 2. Evaluation results of unconditional generation. (A) Distribution of top 3 sc-TM. The white diamonds denote the median value of each group and the dashed line shows the designability threshold of sc-TM = 0.5. **(B)** Success rates of tested methods. Success rate is defined by the number of backbones passing the sc-TM threshold divided by the total number of backbones within each group, with top 3 sc-TM by solid line and best sc-TM by dashed line. The text of numbers refers to the solid lines. **(C)** Distribution of pdb-TM values. The orange half box denotes the distribution and the area filled with dark blue shows the density estimation of values. **(D)** Diversity performances. Diversity is defined as the number of unique clusters divided by the total number of proteins within each group using Foldseek-Cluster. **(E)** Differences between novelty calculation by whether adding the penalty term of designability. Stars on each data group indicate the differences between two types of data within the group, calculated using the T-test. *, p-value < 0.05; **, p-value < 0.01; ***, p-value < 0.001. **(F)** Time usage for generating protein backbones among different methods. The efficiency is calculated as the average time usage per backbone for generating 100 backbones. All tests were conducted under identical hardware resources.

Diversity. We conducted a diversity analysis upon different methods. We used the latest structure clustering algorithm Foldseek-Cluster to process the clustering step using a TM-score threshold of 0.5, where proteins with TM-score above this threshold would be clustered into the same group. The final diversity metric is defined as the number of unique clusters divided by the total number of proteins within each group, with a range from 0 to 1. The results are shown in **Figure 2D**.

Among evaluated methods, *FrameFlow* performed the lowest diversity compared with others, which probably owes to the diversity-designability trade-off mentioned in Yim et al.[53]. *RFdiffusion* and *Genie* showed a continuously increasing diversity by length within their length scope, where *RFdiffusion* started on a low point and reach almost a complete diverse performance when the length of protein is above 800. Interestingly, we also noticed that *FrameDiff* and *Chroma* shows a non-monotonic behavior, with *FrameDiff* declining at the beginning and climbing up later, while *Chroma* behaves in contrast. Generally, the diversity turns higher as the length of protein increases, where the redundancy of training set at different length levels can be a potential contributor.

We also noticed that Foldseek-Cluster failed to perform clustering 50-residue proteins generated by *RFdiffusion*. We speculated the reason to be the low diversity of this group, which made it difficult for Foldseek-Cluster to detect non-redundant secondary structures. To prove this hypothesis, we performed a secondary structure analysis to this group individually, which is shown in **Supplementary Figure 5**. The plot shows that *RFdiffusion* tends to generate protein backbones consisting solely α-helices with a helix percentage range from 80~100%, which partially corroborated our conjecture.

Efficiency. The result displayed in **Figure 2F** demonstrate significant differences in efficiency among various methods. The tested methods showcase most variances when generating short proteins, as *Genie* took about 8 minutes whereas *FrameFlow* took just around 0.15 minutes to generate a backbone, where the latter exhibiting potential for high-throughput applications in conjunction with downstream uses. We also noticed that the growing trends of *Chroma* stabilized when generating long proteins which benefit from their scaling technique[55], which is of great help when generating huge proteins. It is noteworthy that the speeds of current methods are still at the scale of minutes, remains a gap between advanced sequence design methods that generating sequences within a few seconds. Improving inference speed and modeling efficiency may represent an important direction for future development in this field.

Analysis of Structural Properties

Secondary structure analysis. Figure 3A illustrates the SSEs composition from the perspectives of overall average. We first used DSSP to calculate the secondary structures of all generated backbones in a classical 8-string way and converted it to a 4-string style as α-helix, β-sheet, turn and coil. As turns often occur under the circumstances when the protein chain needs to change directions, especially abundant in β-sheets, we suggest that turns and coils should be treated distinctly during analysis, with the former being considered as part of the structured regions. As can be seen from Figure 6A, the average frequency of SSEs by different methods are all more inclined to generate structured proteins, with a α-helix percentage > 50% among most cases. We also noticed that this structured-preference did not change a lot even with the length increased up to 1000. The prefilter step during training where most loop-abundant structures were excluded might be key factor to this behavior.

As the overall average value of SSEs might not reflect the secondary structure of proteins faithfully at an individual level, we also analyzed the density distribution of SSEs of individual proteins, whose results are shown in **Figure 3B**. The distribution of short protein group is relatively broader than the others, where *RFdiffusion* generated helix-mainly proteins. But as length increases, *RFdiffusion* tends to generate proteins with more diverse SSEs than *Chroma*. This distribution combined with the overall suggest that the methods tend to generate proteins mainly with α-helix, which is reasonable not only for the abundant presence of α-helices but also potentially because the short-range interactions are easier for models to learn compared with the long-range interactions in β-sheets and flexible interactions in coils.

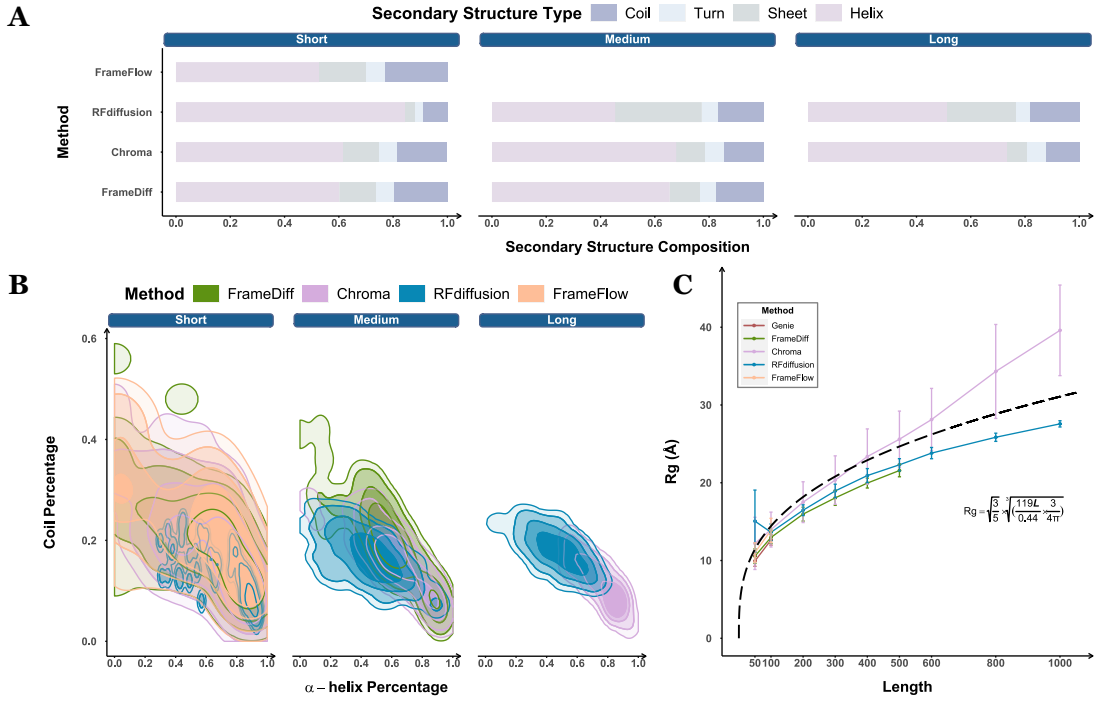

Figure 3. Structural analysis on generated proteins. (A) Average secondary structure distribution within each group. **(B)** Two-dimensional density plot of secondary structure composition of each protein. All data was directly calculated onto protein backbones without refolding. *Genie* was excluded from this test as it solely generates $C\alpha$-only backbones. **(C)** Radius of gyration (R_g) values of different methods. The points are the mean value of R_g and the error bars denote the standard deviation interval within each group. The empirical regression curve of R_g for structured proteins is shown in a dashed style with the text denoted the formula based on previous studies[73, 74].

Radius of gyration. Apart from secondary structures, Rg was also calculated due to its importance to reveal the compactness of protein, whose results are shown in **Figure 3C**. With the empirical regression curve of Rg for structured proteins as a reference, it is shown that most of the methods tends to generate proteins with Rg below the value of corresponding lengths, exhibiting the compactness of generated proteins. *Chroma* is the only one exception with more stretched backbones with more lengths, indicating a potential to design more flexible protein structures.

Ranking of Different Methods

After the analysis of generated protein backbones, we use a ranking procedure based on a procedure combined with TOPSIS decision-making and CRITIC weighting-selection technique, whose result is shown in **Table 3**. We ranked the methods based on the length grouping of short, medium and long, respectively. Based on the ranking

outcome, *FrameFlow* and *Genie* performs the best on short proteins, while *RFdiffusion* shows an outstanding performance when generating long proteins. We also noticed that no single method consistently performs best or worst across all groups, reflecting significant different in the scenarios each method excels in. Consequently, we recommend a flexible and personalized selection of methods in practical applications, tailored to achieve the most optimal performance under specific circumstances.

Table 3. Ranking of Different Methods. Four metrics with arrows are shown as mean values within each group without data preprocessing. C_i denotes the calculated relative nearness value of TOPSIS results, where higher indicates better performances.

Type	Method	Length	Designability(↑)	Novelty(↓)	Diversity (↑)	Efficiency (↓)	C_i	Mean C_i	Rank
Short	Chroma	50	0.691	0.715	0.54	1.17	0.327	0.365	3
		100	0.763	0.682	0.66	2.73	0.402		
	FrameDiff	50	0.781	0.731	0.49	0.44	0.34	0.331	4
		100	0.799	0.711	0.46	1.71	0.321		
	FrameFlow	50	0.807	0.812	0.32	0.11	0.504	0.508	1
		100	0.862	0.721	0.4	0.17	0.512		
	Genie	50	0.841	0.715	0.59	7.92	0.335	0.432	2
		100	0.836	0.607	0.83	8.08	0.529		
	RFdiffusion	50	0.726	0.864	0.35	0.54	0.125	0.235	5
		100	0.956	0.707	0.35	1.05	0.345		
Medium	Chroma	200	0.76	0.676	0.48	4.68	0.275	0.296	3
		300	0.684	0.637	0.67	5.73	0.398		
		400	0.545	0.668	0.74	6.92	0.262		
		500	0.48	0.685	0.84	8.56	0.249		
	FrameDiff	200	0.694	0.666	0.39	1.82	0.524	0.399	2
		300	0.682	0.656	0.48	2.38	0.486		
		400	0.5	0.654	0.55	3.27	0.378		
		500	0.448	0.69	0.59	4.7	0.209		
	RFdiffusion	200	0.952	0.65	0.63	2.54	0.558	0.531	1
		300	0.884	0.614	0.67	6.32	0.51		
		400	0.827	0.591	0.76	10.3	0.541		
		500	0.722	0.596	0.87	14	0.515		
Long	Chroma	600	0.432	0.733	0.88	12.2	0.516	0.354	2
		800	0.434	0.746	0.84	19.27	0.325		
		1000	0.427	0.742	0.7	24.67	0.222		
	RFdiffusion	600	0.654	0.609	0.891	20.26	0.689	0.406	1
		800	0.465	0.677	1	37.44	0.351		
		1000	0.384	0.738	1	54.43	0.178		

Conditional Generation

Apart from the more conceptual unconditional generation, we extended our analysis to motif-scaffolding as a representative problem of conditional generation. An comprehensive experiment along the 24 cases described above was done. Similar to unconditional task, we define a success backbone with refolding structures a sc-TM > 0.5 and motif-RMSD < 1.0Å, with the number of sequences selected by "top N" strategy, where we set $N = 3$ in this case.

Figure 4A displays the overall performance on designability of four tested methods among 24 cases. We observe that *GPDL-H* along with *RFdiffusion* exhibited strong efficacies on most of the cases, while the performance of *Chroma* and *TDS* had a high variance between cases. We also notice that the difficulty to design differs from case by case for all tested methods. For example, *1PRW*, *2KL8* and *3IXT* are designed pretty well by all methods, yet *4JHW*, *5WN9* and *5YUI* are rather difficult to be designed. We also see some interesting properties for different methods, like *Chroma* and *RFdiffusion* tended to behave better to design longer scaffolds given the same motif where *TDS* showed a slightly opposite trend where *GPDL-H* demonstrated the highest robustness among all tested methods. Strikingly, *Chroma* performs beyond expectation as it has not undergone additional fine-tuning for the task of motif-scaffolding, yielding a great potential in designing proteins conditionally within its framework.

Since most of the protein backbones after motif-scaffolding have residue lengths under 200, we suggest that the current definition of backbone success by sc-RMSD < 2Å is also applicable to some extent. We include this result for comparison from in **Supplementary Figure 6**.

Figure 4B showcases the diversity performance of different methods. Similar to the evaluation of designability, clustering was performed within each of the 24 tasks by different methods, where the vacant groups are the ones that Foldseek-Cluster failed to cluster. Based on the outcomes, the performances of diversity exhibited significant variation across different cases, where certain cases such as *5IUS* and *2KL8* displaying markedly low diversity, indicating that all tested methods were capable of generating solely one some specific topology or folding pattern for them. We suggest this might also arises from the proportion of motifs and scaffolds. Due to the incompleteness of result, we did not perform a complete comparison and ranking. However, it is observable that the diversity of GPDL and TDS is marginally superior to the other two

methods.

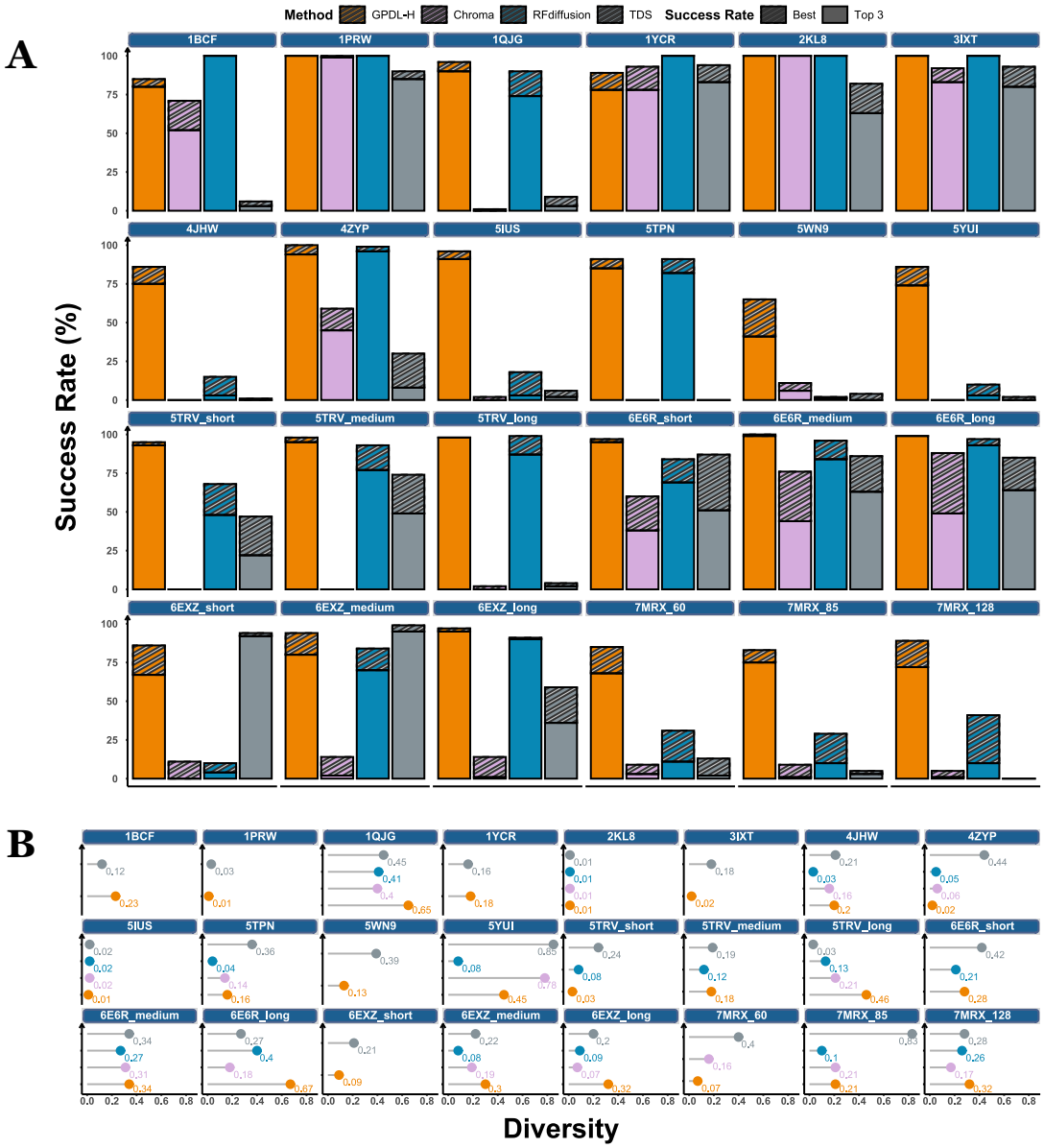

Figure 4. Motif-scaffolding performances for tested methods. Each facet places the result of one specific cases of four methods. **(A)** Success Rates of different methods on designability. The success threshold is defined as refolded sc-TM of the whole structure > 0.5 and motif-RMSD < 1.0Å. The definition of success rate follows the ones with unconditional generation, with the areas with pure color denote the top 3 success rates and the shadow areas as the value surpassed by best success rates. **(B)** Diversity of different methods. The lollipops display the diversity values of sdifferent methods, where missing groups are the ones that Foldseek-Cluster incapable to cluster.

Discussion

In this work, we conducted a thorough study among different protein backbone generation across different types of tasks, offering a precise and comprehensive evaluation of the existing methods. Beyond the qualitative rankings and comparative analyses, it is worthwhile to acknowledge that diversity in short proteins and designability in long proteins represent two primary challenges across different methods. Conversely, given that *novelty* is beside the point to practical biological applications and the lack of substantial disparities it exhibited among the methods, we suggest that its significance may have been overstated in prior studies.

Designing extremely short and large proteins holds particularly utility in specific contexts, such as the former conductive for some peptide drugs[75] or binders[51, 76, 77] and the latter for large multi-domain proteins[78] or symmetrical protein assemblies[79, 80]. While a number of peptide[81] and binder design[82] methods have been proposed to address these needs, the design of large proteins remains an enduring challenge. Moreover, methods focused on generating protein complexes or dynamic backbones, which are key to the design of enzymes or allosteric proteins, are still in a nascent stage. As training data of these methods is rare than protein monomers, there remains a need for either sampling-based methods, such as Markov Chain Monte Carlo (MCMC) sampling[46, 83] and constrained hallucination[23], or for approaches capable of discerning the elusive patterns inherent in protein structures.

Our evaluation process also highlights the shortcomings of current assessment metrics on protein backbone generation methods. To this end, more all-encompassing *in silico* metrics are yet to be developed for both protein backbone generation methods and the quality of individual proteins. The refolding pipeline is also a general framework that different modules can be replaced or parallelly introduced flexibly. As a result, more powerful sequence designers and structure predictors related to these fields will also make significant contributions to this domain.

Given that our evaluation showcases the performance of the current outstanding methods, it also paves the way for future directions in this field. The design targeting long proteins is still in a nascent stage which urgently requires enhancement. The analysis on structural properties reveals a strong preference in structured and helices-abundant proteins, preventing them to generate much coil regions and intrinsically disordered proteins. These proteins are currently studied through molecular dynamics

(MD) simulation[84-86] or conformational sampling[87], which are discarded from most hand-picked training sets of protein backbone generation methods aiming to improve the model fitness for three-dimensional protein structures. With the rapid advancement in this field, more sophisticated tasks in protein backbone generation are coming into view. An example of these tasks is designing of non-idealized proteins attempting to deal with the aforementioned preferences of current methods, which has seen rudimentary progress in [88]. This also indicates the out-of-distribution generalization capability for current methods still has room for improvement.

As for the case of conditional generation, we suggest from our results that this field is still in an early era where remains a gap to bridge before reaching practical downstream applications. Numerous sequence design methods have been developed to tackle the conditional generation tasks from secondary structures to motif-scaffolding, while flow-based models[54] have already shown a potential in both designability and efficiency. Despite of the exceptional capability of powerful generated models, we argue from the outstanding performance of *GPDL-H* that model-based methods like[89-91] still hold relevance in this field. Considering the high-throughput virtual backbone generation demand for downstream tasks, the improvement of efficiency will also be one of the crucial directions for this field.

Even though the comprehensive evaluation conducted in this study, our efforts still require improvements in certain aspects. The refolding pipeline was chosen mainly out of efficiency, where we believe that AlphaFold2 or ColabFold may help increase the precision through the refolding pipeline. The parameter choices of our "top N strategy" and the ranking method we chose could also be potentially either be adapted or complemented to better reveal the robustness of backbones and corresponding methods. Due to the scope of our work, we have not tested more works within the scope of conditional generation, which encompasses a diverse set of tasks and is needed to be evaluated in detail. Some important experimental parameters like noise scale of diffusion models are also interesting factors that influence the performance of models. We intend to periodically update this benchmark by incorporating the latest superior methods, with the objective of contributing to the exhilarating advancements in this field.

Acknowledgements

We would like to thank Xiaochen Cui and Xiaoyue Ji for insightful comments. We thank all past and present members in Chen Lab for helpful feedbacks. We additionally thank Bozitao Zhong for critical discussions upon the manuscript. 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 and 2023YFF1205102), the National Natural Science Foundation of China (21977068 and 32171242), and the Fundamental Research Funds for the Central Universities (YG2023LC03).

References

1. Anfinsen, C.B., *Principles that Govern the Folding of Protein Chains.* Science, 1973. **181**(4096): p. 223-230.
2. Dill, K., R.L. Jernigan, and I. Bahar, *Protein actions: Principles and modeling*. 2017: Garland Science.
3. Huang, P.S., S.E. Boyken, and D. Baker, *The coming of age of de novo protein design.* Nature, 2016. **537**(7620): p. 320-7.
4. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
5. Casadevall, G., C. Duran, and S. Osuna, *AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design.* JACS Au, 2023. **3**(6): p. 1554-1562.
6. Khakzad, H., et al., *A new age in protein design empowered by deep learning.* Cell Systems, 2023. **14**(11): p. 925-939.
7. Kuhlman, B., et al., *Design of a Novel Globular Protein Fold with Atomic-Level Accuracy.* Science, 2003. **302**(5649): p. 1364-1368.
8. Mu, J., et al., *De novo Protein Sequence Design Based on Deep Learning and Validation on CalB Hydrolase.* bioRxiv, 2023: p. 2023.08. 01.551444.
9. Ferruz, N., S. Schmidt, and B. Hocker, *ProtGPT2 is a deep unsupervised language model for protein design.* Nat Commun, 2022. **13**(1): p. 4348.
10. Anand, N., et al., *Protein sequence design with a learned potential.* Nat Commun, 2022. **13**(1): p. 746.
11. Liu, Y., et al., *Rotamer-free protein sequence design based on deep learning and self-consistency.* Nat Comput Sci, 2022. **2**(7): p. 451-462.
12. Dauparas, J., et al., *Robust deep learning–based protein sequence design using ProteinMPNN.* Science, 2022. **378**(6615): p. 49-56.
13. Zheng, Z., et al., *Structure-informed language models are protein designers.* bioRxiv, 2023: p. 2023.02. 03.526917.
14. Ingraham, J., et al., *Generative models for graph-based protein design.* Advances in neural information processing systems, 2019. **32**.
15. Jing, B., et al., *Learning from protein structure with geometric vector perceptrons.* arXiv

preprint arXiv:2009.01411, 2020.
16. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
17. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nat Methods, 2022. **19**(6): p. 679-682.
18. Senior, A.W., et al., *Improved protein structure prediction using potentials from deep learning.* Nature, 2020. **577**(7792): p. 706-710.
19. Zhong, B., et al., *ParaFold: Paralleling AlphaFold for Large-Scale Predictions*, in *International Conference on High Performance Computing in Asia-Pacific Region Workshops*. 2022. p. 1-9.
20. Baek, M., et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
21. Wang, W., Z. Peng, and J. Yang, *Single-sequence protein structure prediction using supervised transformer protein language models.* Nature Computational Science, 2022. **2**(12): p. 804-814.
22. Ovchinnikov, S. and P.S. Huang, *Structure-based protein design with deep learning.* Curr Opin Chem Biol, 2021. **65**: p. 136-144.
23. Anishchenko, I., et al., *De novo protein design by deep network hallucination.* Nature, 2021. **600**(7889): p. 547-552.
24. Huang, B., et al., *A backbone-centred energy function of neural networks for protein design.* Nature, 2022. **602**(7897): p. 523-528.
25. Ho, J., A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models.* Advances in neural information processing systems, 2020. **33**: p. 6840-6851.
26. Song, Y., et al., *Score-based generative modeling through stochastic differential equations.* arXiv preprint arXiv:2011.13456, 2020.
27. Guo, Z., et al., *Diffusion models in bioinformatics and computational biology.* Nature Reviews Bioengineering, 2023: p. 1-19.
28. Ferruz, N., et al., *From sequence to function through structure: Deep learning for protein design.* Comput Struct Biotechnol J, 2023. **21**: p. 238-250.
29. Anand, N. and T. Achim, *Protein structure and sequence generation with equivariant denoising diffusion probabilistic models.* arXiv preprint arXiv:2205.15019, 2022.
30. Lee, J.S., J. Kim, and P.M. Kim, *Score-based generative modeling for de novo protein design.* Nature Computational Science, 2023. **3**(5): p. 382-392.
31. Chu, A.E., et al., *An all-atom protein generative model.* bioRxiv, 2023: p. 2023.05.24.542194.
32. Ni, B., D.L. Kaplan, and M.J. Buehler, *ForceGen: End-to-end de novo protein generation based on nonlinear mechanical unfolding responses using a protein language diffusion model.* arXiv preprint arXiv:2310.10605, 2023.
33. Trippe, B.L., et al., *Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem.* arXiv preprint arXiv:2206.04119, 2022.
34. Watson, J.L., et al., *De novo design of protein structure and function with RFdiffusion.* Nature, 2023. **620**(7976): p. 1089-1100.
35. Castorina, L.V., et al., *PDBench: evaluating computational methods for protein-sequence design. LID - 10.1093/bioinformatics/btad027 [doi] LID - btad027.* (1367-4811

(Electronic)).
36. Gao, Z., et al. *Proteininvbench: Benchmarking protein inverse folding on diverse tasks, models, and metrics*. in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2023.
37. Wang, C., et al. *PDB-Struct: A Comprehensive Benchmark for Structure-based Protein Design*. in *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*. 2023.
38. Yu, J., et al., *Multi-indicator comparative evaluation for deep Learning-Based protein sequence design methods.* Bioinformatics, 2024.
39. Notin, P., et al., *ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction.* bioRxiv, 2023.
40. Buttenschoen, M., G.M. Morris, and C.M. Deane, *PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences.* Chemical Science, 2024.
41. Jiang, L., et al., *De Novo Computational Design of Retro-Aldol Enzymes.* Science, 2008. **319**(5868): p. 1387-1391.
42. Torres, S.V., et al., *De novo design of high-affinity binders of bioactive helical peptides.* Nature, 2023.
43. Procko, E., et al., *A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells.* (1097-4172 (Electronic)).
44. Correia, B.E., et al., *Proof of principle for epitope-focused vaccine design.* Nature, 2014. **507**(7491): p. 201-206.
45. Tischer, D., et al., 2020.
46. Wang, J., et al., *Scaffolding protein functional sites using deep learning.* Science, 2022. **377**(6604): p. 387-394.
47. Wu, L., et al., *Practical and asymptotically exact conditional sampling in diffusion models.* arXiv preprint arXiv:2306.17775, 2023.
48. Alamdari, S., et al., 2023.
49. Tischer, D., et al., *Design of proteins presenting discontinuous functional sites using deep learning.* bioRxiv, 2020: p. 2020.11.29.402743.
50. Zhang, B., et al., *Protein Language Model Supervised Precise and Efficient Protein Backbone Design Method.* bioRxiv, 2023: p. 2023.10.26.564121.
51. Alamdari, S., et al., *Protein generation with evolutionary diffusion: sequence is all you need.* bioRxiv, 2023: p. 2023.09.11.556673.
52. Lin, Y. and M. AlQuraishi, *Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds.* arXiv preprint arXiv:2301.12485, 2023.
53. Yim, J., et al., *SE (3) diffusion model with application to protein backbone generation.* arXiv preprint arXiv:2302.02277, 2023.
54. Yim, J., et al., *Fast protein backbone generation with SE (3) flow matching.* arXiv preprint arXiv:2310.05297, 2023.
55. Ingraham, J.B., et al., *Illuminating protein space with a programmable generative model.* Nature, 2023. **623**(7989): p. 1070-1078.
56. Zhang, Y.e.a. *List of papers about Proteins Design using Deep Learning*. 2021 2024 [cited 2021; Available from:

https://github.com/Peldom/papers_for_protein_design_using_DL.
57. Glasgow, A., et al., *Engineered ACE2 receptor traps potently neutralize SARS-CoV-2.* Proceedings of the National Academy of Sciences, 2020. **117**(45): p. 28046-28055.
58. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins: Structure, Function, and Bioinformatics, 2004. **57**(4): p. 702-710.
59. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score= 0.5?* Bioinformatics, 2010. **26**(7): p. 889-895.
60. Burley, S.K., et al., *RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning.* Nucleic Acids Research, 2022. **51**(D1): p. D488-D508.
61. Berman, H., K. Henrick, and H. Nakamura, *Announcing the worldwide Protein Data Bank.* Nature Structural & Molecular Biology, 2003. **10**(12): p. 980-980.
62. van Kempen, M., et al., *Fast and accurate protein structure search with Foldseek.* Nat Biotechnol, 2023.
63. Barrio-Hernandez, I., et al., *Clustering predicted structures at the scale of the known protein universe.* Nature, 2023. **622**(7983): p. 637-645.
64. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers: Original Research on Biomolecules, 1983. **22**(12): p. 2577-2637.
65. Joosten, R.P., et al., *A series of PDB related databases for everyday needs.* Nucleic acids research, 2010. **39**(suppl_1): p. D411-D419.
66. Kunzmann, P. and K. Hamacher, *Biotite: a unifying open source computational biology framework in Python.* BMC Bioinformatics, 2018. **19**(1): p. 346.
67. Mu, J., et al., *Recent force field strategies for intrinsically disordered proteins.* Journal of chemical information and modeling, 2021. **61**(3): p. 1037-1047.
68. Tzeng, G.-H. and J.-J. Huang, *Multiple attribute decision making: methods and applications*. 2011: CRC press.
69. Saaty, T.L., *What is the analytic hierarchy process?* 1988: Springer.
70. Diakoulaki, D., G. Mavrotas, and L. Papayannakis, *Determining objective weights in multiple criteria problems: The critic method.* Computers & Operations Research, 1995. **22**(7): p. 763-770.
71. Mardani, A., A. Jusoh, and E.K. Zavadskas, *Fuzzy multiple criteria decision-making techniques and applications – Two decades review from 1994 to 2014.* Expert Systems with Applications, 2015. **42**(8): p. 4126-4148.
72. McKnight, P.E. and J. Najab, *Mann-Whitney U Test*, in *The Corsini Encyclopedia of Psychology*. p. 1-1.
73. Xiang, S., et al., *Phosphorylation Drives a Dynamic Switch in Serine/Arginine-Rich Proteins.* Structure, 2013. **21**(12): p. 2162-2174.
74. Sterckx, Yann G.J., et al., *Small-Angle X-Ray Scattering- and Nuclear Magnetic Resonance-Derived Conformational Ensemble of the Highly Flexible Antitoxin PaaA2.* Structure, 2014. **22**(6): p. 854-865.
75. Chang, L., et al., *Revolutionizing peptide-based drug discovery: Advances in the post-*

AlphaFold era. WIREs Computational Molecular Science, 2024. **14**(1): p. e1693.
76. Bennett, N.R., et al., *Improving de novo protein binder design with deep learning.* Nature Communications, 2023. **14**(1): p. 2625.
77. Gainza, P., et al., *De novo design of protein interactions with learned surface fingerprints.* Nature, 2023. **617**(7959): p. 176-184.
78. Khersonsky, O. and S.J. Fleishman, *What Have We Learned from Design of Function in Large Proteins?* BioDesign Research, 2022. **2022**: p. 9787581.
79. Wicky, B.I.M., et al., *Hallucinating symmetric protein assemblies.* Science, 2022. **378**(6615): p. 56-61.
80. and, D.S.G. and A.J. Olson, *Structural Symmetry and Protein Function.* Annual Review of Biophysics and Biomolecular Structure, 2000. **29**(1): p. 105-153.
81. An, L., et al., *De novo design of diverse small molecule binders and sensors using Shape Complementary Pseudocycles.* bioRxiv, 2023: p. 2023.12.20.572602.
82. Rettie, S.A., et al., *Cyclic peptide structure prediction and design using AlphaFold.* bioRxiv, 2023: p. 2023.02.25.529956.
83. Yeh, A.H.-W., et al., *De novo design of luciferases using deep learning.* Nature, 2023. **614**(7949): p. 774-780.
84. Ji, X., et al., *Personal Precise Force Field for Intrinsically Disordered and Ordered Proteins Based on Deep Learning.* Journal of Chemical Information and Modeling, 2023. **63**(1): p. 362-374.
85. Cui, X., H. Liu, and H.-F. Chen, *Polarizable Force Field of Intrinsically Disordered Proteins with CMAP and Reweighting Optimization.* Journal of Chemical Information and Modeling, 2022. **62**(20): p. 4970-4982.
86. Song, G., et al., *Phosphorylation Modification Force Field FB18CMAP Improving Conformation Sampling of Phosphoproteins.* Journal of Chemical Information and Modeling, 2023. **63**(5): p. 1602-1614.
87. Zhu, J., et al., *Phanto-IDP: compact model for precise intrinsically disordered protein backbone generation and enhanced sampling.* Brief Bioinform, 2023. **25**(1).
88. Haiyan, L., L. Yufeng, and C. Linghui, *Diffusion in a quantized vector space generates non-idealized protein structures and predicts conformational distributions.* bioRxiv, 2023: p. 2023.11. 18.567666.
89. Frank, C., et al., *Efficient and scalable de novo protein design using a relaxed sequence space.* bioRxiv, 2023: p. 2023.02.24.529906.
90. Jeliazkov, J.R., D.d. Alamo, and J.D. Karpiak, *ESMFold Hallucinates Native-Like Protein Sequences.* bioRxiv, 2023: p. 2023.05.23.541774.
91. Verkuil, R., et al., *Language models generalize beyond natural proteins.* bioRxiv, 2022: p. 2022.12.21.521521.