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

Zhuoqi Zheng†, Bo Zhang†, Bozitao Zhong, Kexin Liu, Zhengxin Li, Junjie Zhu, Jinyu Yu, 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, 800 Dongchuan Road, 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, especially for backbone generation, which stands out as more challenging yet valuable, offering the ability to design novel protein folds with fewer constraints. However, a comprehensive delineation of its potential for practical application in protein engineering remains lacking, as does 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 like designability, novelty, diversity, efficiency and structural properties. We also extrapolated our benchmark to include the motif-scaffolding problem, demonstrating the utility of these conditional generation models. Our findings reveal that *FrameFlow* and *RFdiffusion* in unconditional generation and *GPDL-H* in conditional generation showcased the most outstanding performances. Furthermore, we described a systematic study to investigate conditional generation and applied it to the motif-scaffolding task, offering a novel perspective for the analysis and development of conditional protein design methods. All data and scripts will be available at https://github.com/Immortals-33/Scaffold-Lab.

Key Words: *de novo* 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. The magnificent breakthroughs in protein design along with protein structure prediction[4] have allowed biologists to dive deeper into the mystery of protein structures, breathing a new life into the field of protein science and protein engineering[5,6].

Protein design has undergone a paradigm shift over the past few decades. In the early stages of *de novo* protein design, researchers used physical-based methods to search through rotamer libraries in order to design new protein sequences based on a pre-defined structure scaffold[7]. In the past couple of years, 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]. Protein backbone generation was carried out to directly manipulate 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 over the recent years, for example, denoising diffusion models[25,26] have 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 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 needed later to find the optimal sequences that folding into the designed backbones. Moreover, since protein backbone generation is targeted for novel structure generation and is an upstream step in protein engineering, there is still a lack of knowledge about how these methods can be applied to real-world

applications. Furthermore, due to the varying metrics and length scopes employed by different methods, a set of benchmark for assessing different protein backbone generation methods is urgently needed, especially compared to the various established works in place for sequence design[35-38], protein fitness prediction[39] and molecular generation tasks[40].

Here 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, then carefully grouped them into different categories according to different lengths of proteins. We divided the benchmark sets into two different problems. 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 pivotal input parameter. Unconditional generation has been a proof-of-concept exploration for almost all protein backbone generation methods as 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 since it is the destination to fully-programmed roadmap of protein design, aiming to develop *de novo* proteins with properties specified by scientists. 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[33,34,45-48]. 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.

Beyond prevalent computational metrics, we also set up a personalized scheme to gain a deeper understanding of the progress and limitations of these methods. We first performed analysis on structural properties for unconditional generation, indicating that current methods exhibit a strong propensity for generating highly structured proteins. In the context of conditional generation, we transcended widely-used computational metrics and delved into the underlying reasons for different outcomes of designs. Furthermore, the capacity to create ideal protein structures is a prerequisite of

successful designs rather than meeting certain conditional criteria. In summary, our research undertook 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*[49] and *FrameDiff*[50] used 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] integrated a diffusion models with a pretrained structure prediction model RoseTTAFold[19] to greatly improve the performance. *Chroma* incorporated 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*[46] and *FrameFlow*[51] was developed upon *FrameDiff*'s architecture, with the former used a particle filtering strategy to raise conditional sampling applied to motif-scaffolding task, and the latter replaced 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.[45], *GPDL*[48] leveraged 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[49]	2023-01	Diffusion-based	4.1 M	Unconditional
FrameDiff[50]	2023-03	Diffusion-based	17.4 M	Unconditional
RFdiffusion[34]	2023-04	Diffusion-based	59.8 M	Both
TDS[46]	2023-06	Diffusion-	17.4 M	Conditional

		based		
GPDL[48]	2023-10	ESM-based	3 B	Conditional
FrameFlow[51]	2023-10	Flow-based	16.8 M	Unconditional*
Chroma[52]	2023-11	Diffusion-based	18.4 M	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.[53].
‡M = million; B = billion.

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 key 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*[54] 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 information is listed in **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 and 10 sequences selected, we obtained 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.

We used ProteinMPNN[12] as sequence designer throughout our pipeline which had been widely utilized in evaluation of protein backbone generation methods[34,49-51] and validated by wet-lab experiments[34,45,55]. Out of consideration of the potential bias by different sequence design methods, we compared the performances ProteinMPNN to another state-of-the-art method, ESM-IF1[56]. The results in **Supplementary Figure 1** indicated that ProteinMPNN is able to find sequences with higher refolding ability than ESM-IF1, hence maximizing the potential of different protein backbone generation methods. For the structure predictor, we compared three prevalent highly accurate protein structure prediction methods, AlphaFold2[4], ESMFold[16] and OmegaFold[57]. The result showcased that three structure prediction methods exhibit a relatively strong correlation (shown in **Supplementary Figure 2)**. Thus, we used ESMFold out of its outstanding performance on *de novo* proteins and its high computational efficiency.

Conditional Generation

For conditional generation, we chose *motif-scaffolding* as the essential problem to be solved in conditional generation, where the it is defined as below: Given a functional "motif" harbored 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.

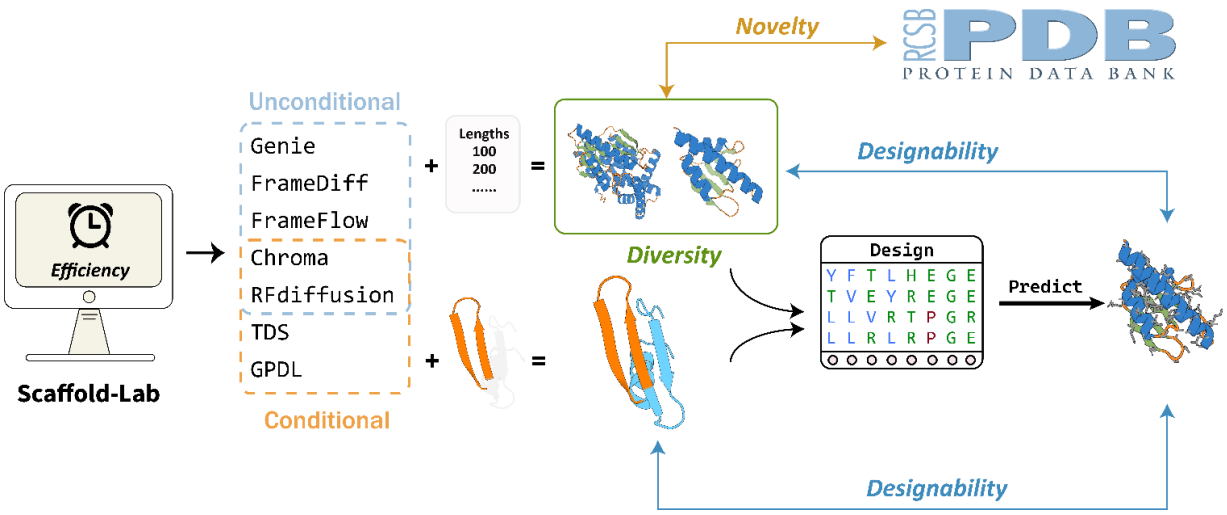

Figure 1. The overall workflow of Scaffold-Lab. Seven state-of-the-art protein backbone generation methods were evaluated by two types of tasks throughout a refolding pipeline. Four qualitative metrics were introduced to qualitatively assess the performances of methods, namely designability, novelty,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

Evaluation Metrics

In this part, we introduced the metrics used in this evaluation, encompassing a range of both quantitative and qualitative aspects. We divided these metrics into five categories, 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, a brief introduction was done for each of these metrics.

Designability

Designability measures the quality of generation, which is the most important metric for protein backbone generation tasks. Designability assesses 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 designable when the refolding structures and original backbones shares a TM-score[58] > 0.5, which are considered to have the same fold pattern[59]. Compared to sc-RMSD, sc-TM could show better consistency through protein with different lengths (shown in **Supplementary Figures 3-4**). We also adopted a "Top N strategy". While current works considered designing solely one sequence for the sc-TM criteria, the number of different sequences that fold into the generated structures could also reveal the robustness and fitness of backbones. Therefore, we reported the sc-TM as the mean value of N sequences with highest sc-TM among generated sequences, which 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 (shown in **Figure 2B** and **Supplementary Figure 5)**.

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

Novelty

Novelty measures the structural difference between the generated protein backbones to all existing proteins. We followed the novelty metric raised by Yim *et al.*[50] by calculating the highest TM-score between generated protein backbones and all proteins in Protein Data Bank (PDB)[60,61], which is referred as *pdb-TM*. If *pdb-TM* is low, it means that the generated backbones are novel compared to most of the existing

experimentally solved native proteins, and vice versa. We used Foldseek[62] to search generated backbones against the PDB database. to find their nearest protein structure. In order to make a fair comparison while maintaining computational efficiency, we applied the same parameter suite to all tested methods.

Apart from pdb-TM, we supposed that there exist certain generated backbones which are novel but undesignable. Thus, the *novelty score* was added a penalty to poorly designable backbones when calculating novelty, formulated as below:

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

In the equation, n refers to the "Top N" strategy when picking sequences generated by ProteinMPNN, 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 defined sc-TM > 0.5 as the threshold of designability, we left the backbones with sc-TM > 0.5 as original pdbTM value, otherwise a penalty is put on the shortfall part between sc-TM and the passing threshold, 0.5. After the reconsideration of novelty metric, the discrepancy between designable novel backbones and non-designable novel backbones would be exposed to a significant extent.

Diversity

Diversity was designed to measure internal discrepancy of generated samples, 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 could benefit both the heterogeneity of potential application and the success rate for 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 protein lengths and used Foldseek-Cluster[63] to cluster each group of protein backbones. We used a clustering TM-Score threshold of 0.5 based on the aforementioned consensus[59]. For the definition of diversity, we followed Yim *et al.*[50] as the number of unique clusters divided by the total number of protein backbones each group:

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

Generally, 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 maintained 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 necessary to check the physical plausibility and quality from structural perspective. In this part, we analyzed 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 sequence design methods might bring internal bias for SSE preferences[37], we directly computed SSE for the original generated protein backbones. We calculated SSE using DSSP software[64,65] through Biotite[66], count the eight types of SSEs assigned by DSSP and further group them into four common secondary structure types. 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, neutron scattering and other experiments, which is one of the key properties of biological polymer molecule as it measures the size or compactness of proteins. Rg can be a key distinction reflect the flexibility difference between structured proteins and intrinsically disordered proteins (IDPs), 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 compared to an experienced Rg-protein length diagram. The 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[37], 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 introduced 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 used this data preprocessing procedure, and ranked different methods based on the length groups, where the four qualitative metrics are **designability**, **novelty**, **diversity** and **efficiency**. We suggested interested readers to Supplementary Text for a detailed version of our ranking procedure.

Results

Unconditional Generation

Designability. The overall results of unconditional generation task are shown in **Figure**

2. We first reported the top 3 sc-TM distribution of different methods along different groups and shown in **Figure 2A**. From the result of short-protein group, nearly all the evaluated methods performed well, where all the methods attained more than 85% designable backbones. 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. We noted that *FrameFlow*, *Genie* and *FrameDiff* generated short proteins that almost all designable while *RFdiffusion* performed the most robust along all lengths.

To further explore the qualities of generated backbones, we calculated the success rates 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* performed the best among all tested methods, while other methods are also reliable when generating short proteins. 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. The sc-RMSD values also supported the conclusion (**Supplementary Figure 4)**. As a reference, we also reported the best sc-TM distribution, which is shown in **Supplementary Figure 5**.

Novelty. We performed novelty analysis to different length groups, whose results are shown in **Figure 2C**. The pdb-TM values predominantly range between 0.6 and 0.7, with no significant differences in trends observed between various methods and length categories. Notably, *FrameFlow* and *RFdiffusion* exhibit slightly highest pdb-TM values in the group of short proteins, 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 might 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** shows 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 listed in **Supplementary Table 2**.

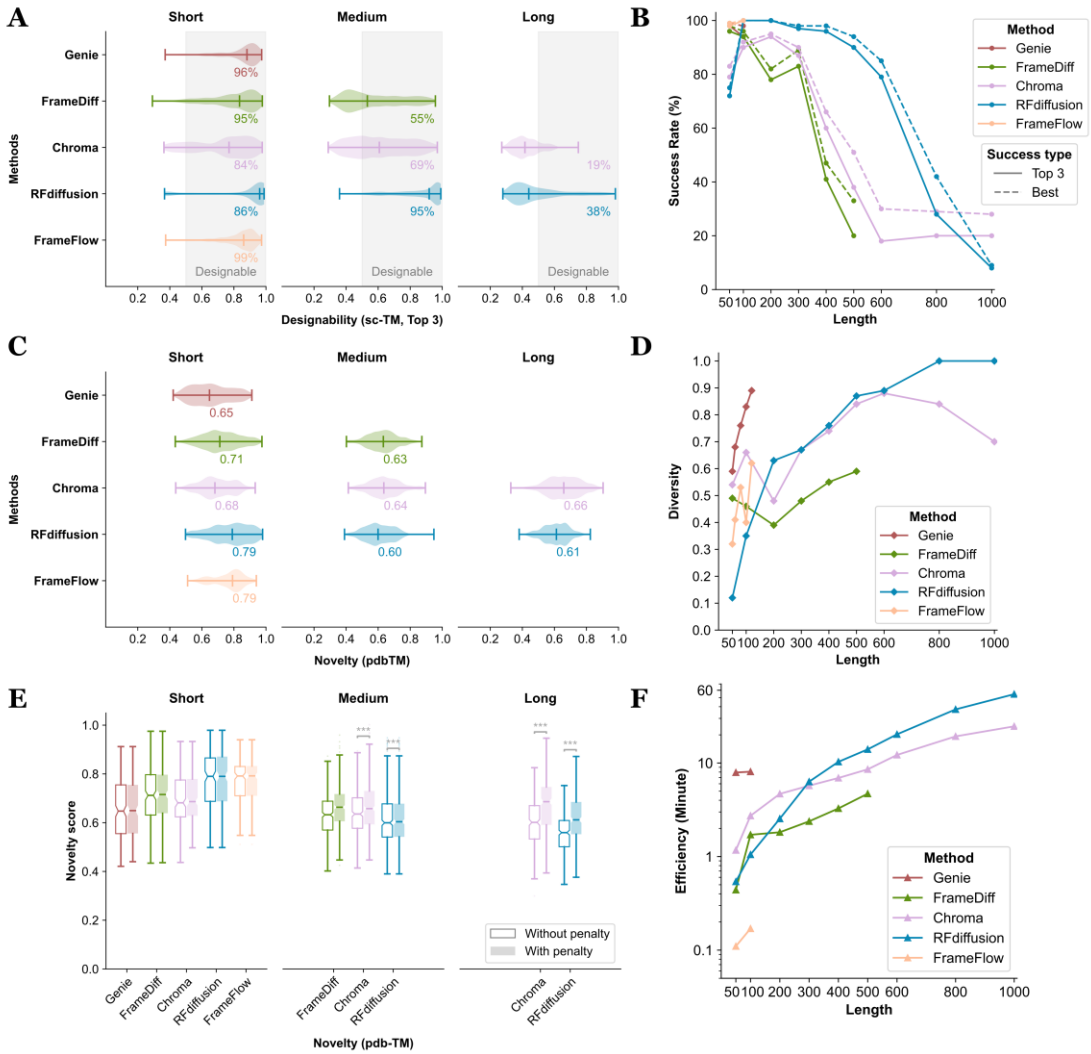

Figure 2. Evaluation results of unconditional generation. (A) Distribution of top 3 sc-TM. The number attached to each group denotes the proportion of backbones with top 3 sc-TM ⩾ 0.5. **(B)** Success rates of tested methods. Success rate is defined by the number of backbones passing the sc-TM threshold 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 number attached to each group denotes the median value. **(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**.

FrameFlow had the lowest diversity compared with other methods, which probably owes to the diversity-designability trade-off mentioned in Yim *et al*.[50]. *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.

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[52], 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. 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. *RFdiffusion* exhibits a strong tendency to generate proteins consisting mostly of α-helices, while it tends to produce proteins with more diverse SSEs than *Chroma* as length increases. This distribution combined with the overall suggest that the methods tend to generate helical proteins, 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.

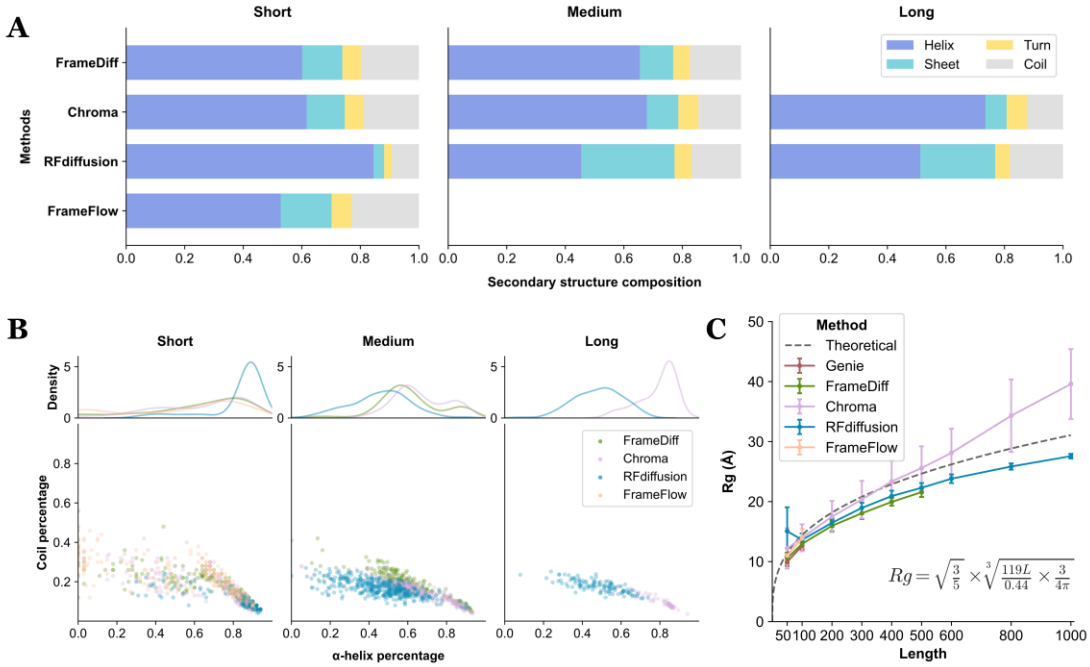

Figure 3. Structural analysis on generated proteins. (A) Average secondary structure distribution within each method. **(B)** Distribution of secondary structure composition of each protein. All data was curated by directly calculating onto protein backbones without refolding. *Genie* was excluded from this test as it solely generates $C\alpha$-only backbones. **(C)** Radius of gyration (Rg) values of different methods. The points are the mean value of Rg and the error bars denote the standard deviation interval within each group. The empirical regression curve of Rg 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 gathered 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 upon 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	0.368	**0.403**	**3**
		100	0.763	0.682	0.66	2.73	0.437		
	FrameDiff	50	0.781	0.731	0.49	0.44	0.378	**0.365**	**4**
		100	0.799	0.711	0.46	1.71	0.353		
	FrameFlow	50	0.807	0.812	0.32	0.11	0.509	**0.517**	**1**
		100	0.862	0.721	0.4	0.17	0.525		
	Genie	50	0.841	0.715	0.59	7.92	0.377	**0.464**	**2**
		100	0.836	0.607	0.83	8.08	0.551		
	RFdiffusion	50	0.726	0.864	0.12	0.54	0.116	**0.238**	**5**
		100	0.956	0.707	0.35	1.05	0.36		
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		

	600	0.654	0.609	0.891	20.26	0.689		
RFdiffusion	800	0.465	0.677	1	37.44	0.351	**0.406**	**1**
	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. We conducted a comprehensive experiment across the 24 cases described above. Similar to the unconditional task, we defined a success backbone with refolding structures a sc-TM > 0.5 and motif-RMSD < 1.0 Å, with the number of sequences selected by "top N" strategy, where we set $N = 3$ in this case.

Figure 4A displays the average performance on designability of four tested methods along 24 cases. We observed that *GPDL-H* and *RFdiffusion* exhibited strong efficacies with an average success rate over 50%, during which *GPDL-H* also generated considerably diverse backbones. By way of comparison, *Chroma* and *TDS* had modest performances on success rates while showcased high variances of distribution, implicating that they also performed well on some certain cases. To explore the differences of success rate and diversity between distinct cases, we computed these metrics for each individual case (shown in **Supplementary Figure 6**). The results show that different methods have their particular strengths in designing certain cases, while *GPDL-H* showed a relatively robust performance among all cases. We inspected that the cases can be loosely divided to three categories: *Simple*, *Moderate* and *Difficult*. *Simple* cases can be designed quite well along all tested method, such as *2KL8* (**Figure 4B**), while *Moderate* cases show diverging performances by different methods such as *4JHW* (**Figure 4C**). *Difficult* cases, like *5WN9* **(Figure 4D)**, prove challenging to all methods. We argue that this diverging performance potentially result from different secondary structures and topologies of certain motifs as well as the length scopes of motifs and scaffolds for different cases. For instance, tackling *2KL8* needs designing a 20-residue scaffold around a 59-residue fixed motif, where the predefined well-structuralized motif allows models to sculpt a small scaffold more easily. To make a comparison between different success definition, we also calculated the individual success rate under the definition using sc-RMSD < 2.0 Å instead of sc-TM > 0.5 (shown in **Supplementary Figure 7**), which showcases a similar trend.

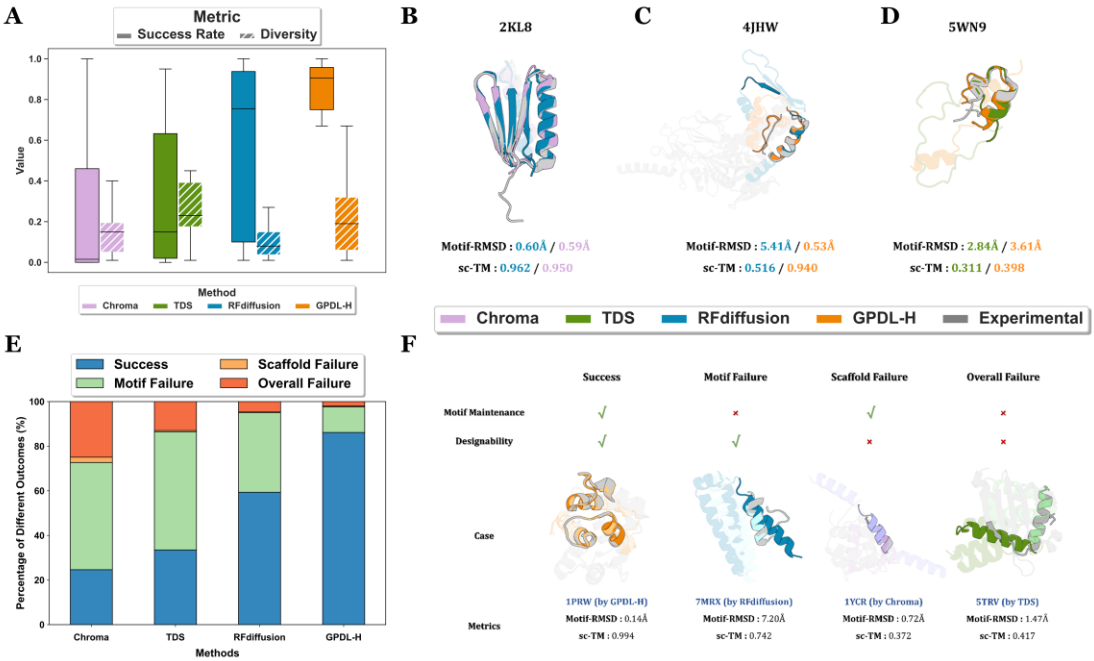

Figure 4. Motif-scaffolding results and failure mode analysis for tested methods. (A) Overall distribution of success rate and diversity across the 24-cases benchmark. The success definition follows the "Top N" strategy in unconditional generation. **(B)**, **(C)**, **(D)** The difficulty of solving motif-scaffolding problems varies across different cases. **(B)** An illustration of *Simple* cases represented by *2KL8*. **(C)** An illustration of *Moderate* cases represented by *4JHW*. **(D)** An illustration of *Difficult* cases represented by *5WN9*. In **(B)**, **(C)** and **(D)**, refolded structures of designs by different methods (shown in the legend on top of Figure 4F) are overlayed with crystal structures from PDB (shown in gray). Motifs are displayed as *opaque* and scaffolds are displayed as *transparent*. **(E)** Percentage of different outcomes of motif-scaffolding task among the 24-cases dataset. Each color denotes a unique type of outcome. **(F)** Schematic diagram of different outcomes. Each case in the "case" panel denotes a unique type of outcome and different colors denote different methods (shown in the legend on top of Figure 4F). Inside each scheme, we use *dark* colors to represent *refolded structures* after folding and *light* colors to denote *original backbones* designed by different methods. Experimental structures are shown in gray, motifs are displayed as *opaque* and scaffolds are displayed as *transparent*. The motif-RMSD is calculated between refolded structures and experimental structures from PDB.

We intended to gain a deeper understanding on the potential causes of unsuccessful designs beyond the explicit performances of different methods. Thus, we performed an analysis on the different outcomes of each motif-scaffolding task by grouping them into four categories. Since the success criterion by satisfying both motif-RMSD < 1.0 Å and sc-TM > 0.5 corresponds separately to motif maintenance and overall designability, we further characterized the failure cases into three types: We defined the designs that fulfill motif maintenance but not overall designability as *motif failure* and the opposite ones as *scaffold failure*, plus the ones that failed to satisfy either criterion as *overall*

failure. A failure mode analysis was conducted based on this classification which is shown in **Figure 4E**. It could be easily observed that *motif failure* dominated the failed cases, followed by *overall failure*. It is noteworthy that *scaffold failure* accounts for a minimal fraction of failed cases, indicating that meeting the overall designability is a necessary but not sufficient condition for a successful design.

The analysis on failure modes offers us a chance to dive deeper into the nature of conditional generation. The fact that overall designability is a prerequisite for solving the motif-scaffolding problems emphasizes the necessity of learning a reasonable protein structure compared to meeting a certain substructure. Since the motif-scaffolding problem is a representative task of conditional generation, we propose that this mode can be generalized to the broader domain of conditional generation. In other words, designing a physically-plausible protein structure is more essential than fulfilling certain conditions. Based on this concept, we supposed that extending a well-established unconditional generation model to conditional generation by certain modifications or fine-tuning is more efficient than developing a conditional generation model from scratch, as the former strategy stems from a model that has already captured the underlying distributions of protein structures to a certain extent.

Discussion

In this work, we conducted a thorough study of different protein backbone generation methods across various types of tasks, providing a precise and comprehensive evaluation. 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 mini-binders[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 for the case of conditional generation, we suggest from the results that this field is still in an early era where there remains a gap to bridge before reaching practical downstream applications. The fact that *motif failure* accounts the most of failures indicates that meeting certain conditions are still challenging for these models. While several sequence design methods have been developed to tackle the task of conditional generation[83,84], protein backbone generation methods still hold considerable potential for enhancement. The direct modeling of protein structures stands a powerful approach to manipulate the functions and desired properties of designed proteins. We propose that both clearer definitions of numerous conditional generation problems and customized training strategies like amortization would be instrumental in improving the capacities of these methods. Despite the exceptional capability of powerful generated models, the outstanding performance of *GPDL-H* demonstrates that model-based methods like[85-87] still hold relevance in this field. Considering the high-throughput virtual backbone generation demand for downstream tasks, the improvement of efficiency is another promising avenue for exploration.

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[88-90] or conformational sampling[91], which are discarded from most hand-picked training sets of protein backbone generation methods aiming to improve the fitness of models. An example of these tasks is the design of non-idealized proteins attempting to deal with the aforementioned preferences of current methods, which has seen progress in [92]. We suggest that a higher-level representation for protein structures as well as augmentation in training data would be highly useful for the generalizability of different methods. With the rapid advancement in recent years, we are witnessing an avalanche of sophisticated tasks in protein backbone generation coming into view, offering greater challenges and prospect in protein discoveries.

Acknowledgements

We would like to thank Xiaochen Cui, Xiaoyue Ji and Taeyoung Choi for insightful comments. We thank Tao Guo for helpful discussions on practical usage related to Foldseek-Cluster. This work was supported by the Center for HPC at Shanghai Jiao Tong University, and the National Key Research and Development Program of China (2023YFF1205102 and 2020YFA0907700), the Fundamental Research Funds for the Central Universities (YG2023LC03), and the National Natural Science Foundation of China (21977068 and 32171242).

References

1 Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223-230 (1973). https://doi.org/doi:10.1126/science.181.4096.223

2 Dill, K., Jernigan, R. L. & Bahar, I. *Protein actions: Principles and modeling*. (Garland Science, 2017).

3 Huang, P. S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320-327 (2016). https://doi.org/10.1038/nature19946

4 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org/10.1038/s41586-021-03819-2

5 Casadevall, G., Duran, C. & Osuna, S. AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design. *JACS Au* **3**, 1554-1562 (2023). https://doi.org/10.1021/jacsau.3c00188

6 Khakzad, H. *et al.* A new age in protein design empowered by deep learning. *Cell Systems* **14**, 925-939 (2023). https://doi.org/https://doi.org/10.1016/j.cels.2023.10.006

7 Kuhlman, B. *et al.* Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **302**, 1364-1368 (2003). https://doi.org/doi:10.1126/science.1089427

8 Mu, J. *et al.* De novo Protein Sequence Design Based on Deep Learning and Validation on CalB Hydrolase. *bioRxiv*, 2023.2008. 2001.551444 (2023).

9 Ferruz, N., Schmidt, S. & Hocker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* **13**, 4348 (2022). https://doi.org/10.1038/s41467-022-32007-7

10 Anand, N. *et al.* Protein sequence design with a learned potential. *Nat Commun* **13**, 746 (2022). https://doi.org/10.1038/s41467-022-28313-9

11 Liu, Y. *et al.* Rotamer-free protein sequence design based on deep learning and self-consistency. *Nat Comput Sci* **2**, 451-462 (2022). https://doi.org/10.1038/s43588-022-00273-6

12 Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022). https://doi.org/doi:10.1126/science.add2187

13 Zheng, Z. *et al.* Structure-informed language models are protein designers. *bioRxiv*, 2023.2002. 2003.526917 (2023).

14 Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative models for graph-based

protein design. *Advances in neural information processing systems* **32** (2019).
15 Jing, B., Eismann, S., Suriana, P., Townshend, R. J. & Dror, R. 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* **379**, 1123-1130 (2023). https://doi.org/10.1126/science.ade2574
17 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682 (2022). https://doi.org/10.1038/s41592-022-01488-1
18 Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020). https://doi.org/10.1038/s41586-019-1923-7
19 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021). https://doi.org/10.1126/science.abj8754
20 Wang, W., Peng, Z. & Yang, J. Single-sequence protein structure prediction using supervised transformer protein language models. *Nature Computational Science* **2**, 804-814 (2022). https://doi.org/10.1038/s43588-022-00373-3
21 Zhong, B. *et al.* ParaFold: Paralleling AlphaFold for Large-Scale Predictions. *arXiv preprint arXiv:2111.06340* (2021).
22 Ovchinnikov, S. & Huang, P. S. Structure-based protein design with deep learning. *Curr Opin Chem Biol* **65**, 136-144 (2021). https://doi.org/10.1016/j.cbpa.2021.08.004
23 Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547-552 (2021). https://doi.org/10.1038/s41586-021-04184-w
24 Huang, B. *et al.* A backbone-centred energy function of neural networks for protein design. *Nature* **602**, 523-528 (2022). https://doi.org/10.1038/s41586-021-04383-5
25 Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840-6851 (2020).
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*, 1-19 (2023). https://doi.org/https://doi.org/10.1038/s44222-023-00114-9
28 Ferruz, N. *et al.* From sequence to function through structure: Deep learning for protein design. *Comput Struct Biotechnol J* **21**, 238-250 (2023). https://doi.org/10.1016/j.csbj.2022.11.014
29 Anand, N. & Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019* (2022).
30 Lee, J. S., Kim, J. & Kim, P. M. Score-based generative modeling for de novo protein design. *Nature Computational Science* **3**, 382-392 (2023). https://doi.org/10.1038/s43588-023-00440-3
31 Chu, A. E., Cheng, L., Nesr, G. E., Xu, M. & Huang, P.-S. An all-atom protein generative model. *bioRxiv*, 2023.2005.2024.542194 (2023). https://doi.org/10.1101/2023.05.24.542194
32 Ni, B., Kaplan, D. L. & Buehler, M. J. 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* **620**, 1089-1100 (2023). https://doi.org/10.1038/s41586-023-06415-8
35 Castorina, L. V., Petrenas, R., Subr, K. & Wood, C. A.-O. PDBench: evaluating computational methods for protein-sequence design. **39** (2023). https://doi.org/https://doi.org/10.1093/bioinformatics/btad027
36 Gao, Z. *et al.* in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2023).
37 Yu, J., Mu, J., Wei, T. & Chen, H.-F. Multi-indicator comparative evaluation for deep Learning-Based protein sequence design methods. *Bioinformatics* (2024). https://doi.org/10.1093/bioinformatics/btae037
38 Wang, C. *et al.* PDB-Struct: A Comprehensive Benchmark for Structure-based Protein Design. *arXiv preprint arXiv:2312.00080* (2023).
39 Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. *bioRxiv* (2023). https://doi.org/10.1101/2023.12.07.570727
40 Buttenschoen, M., Morris, G. M. & Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science* (2024). https://doi.org/10.1039/d3sc04185a
41 Jiang, L. *et al.* De Novo Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387-1391 (2008). https://doi.org/doi:10.1126/science.1152692
42 Torres, S. V. *et al.* De novo design of high-affinity binders of bioactive helical peptides. *Nature* (2023). https://doi.org/10.1038/s41586-023-06953-1
43 Procko, E. *et al.* A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. *Cell* **157**, 1644-1656 (2014). https://doi.org/https://doi.org/10.1016/j.cell.2014.04.034
44 Correia, B. E. *et al.* Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201-206 (2014). https://doi.org/10.1038/nature12966
45 Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387-394 (2022). https://doi.org/10.1126/science.abn2100
46 Wu, L., Trippe, B. L., Naesseth, C. A., Blei, D. M. & Cunningham, J. P. Practical and asymptotically exact conditional sampling in diffusion models. *arXiv preprint arXiv:2306.17775* (2023).
47 Alamdari, S. *et al.* Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023.2009.2011.556673 (2023). https://doi.org/10.1101/2023.09.11.556673
48 Zhang, B. *et al.* Protein Language Model Supervised Precise and Efficient Protein Backbone Design Method. *bioRxiv*, 2023.2010.2026.564121 (2023). https://doi.org/10.1101/2023.10.26.564121
49 Lin, Y. & AlQuraishi, M. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *arXiv preprint arXiv:2301.12485* (2023).
50 Yim, J. *et al.* SE (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277* (2023).

51 Yim, J. *et al.* Fast protein backbone generation with SE (3) flow matching. *arXiv preprint arXiv:2310.05297* (2023).
52 Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070-1078 (2023). https://doi.org/10.1038/s41586-023-06728-8
53 Zhang, Y. e. a. *List of papers about Proteins Design using Deep Learning*, <https://github.com/Peldom/papers_for_protein_design_using_DL> (2021).
54 Glasgow, A. *et al.* Engineered ACE2 receptor traps potently neutralize SARS-CoV-2. *Proceedings of the National Academy of Sciences* **117**, 28046-28055 (2020). https://doi.org/doi:10.1073/pnas.2016093117
55 Sumida, K. H. *et al.* Improving Protein Expression, Stability, and Function with ProteinMPNN. *Journal of the American Chemical Society* (2024). https://doi.org/10.1021/jacs.3c10941
56 Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022.2004.2010.487779 (2022). https://doi.org/10.1101/2022.04.10.487779
57 Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022.2007.2021.500999 (2022). https://doi.org/10.1101/2022.07.21.500999
58 Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**, 702-710 (2004). https://doi.org/https://doi.org/10.1002/prot.21643
59 Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics* **26**, 889-895 (2010). https://doi.org/https://doi.org/10.1093/bioinformatics/btq066
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* **51**, D488-D508 (2022). https://doi.org/10.1093/nar/gkac1077
61 Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology* **10**, 980-980 (2003). https://doi.org/10.1038/nsb1203-980
62 van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* (2023). https://doi.org/10.1038/s41587-023-01773-0
63 Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637-645 (2023). https://doi.org/10.1038/s41586-023-06510-w
64 Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* **22**, 2577-2637 (1983). https://doi.org/https://doi.org/10.1002/bip.360221211
65 Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic acids research* **43**, D411–D419 (2010). https://doi.org/https://doi.org/10.1093/nar/gkq1105
66 Kunzmann, P. & Hamacher, K. Biotite: a unifying open source computational biology framework in Python. *BMC Bioinformatics* **19**, 346 (2018). https://doi.org/10.1186/s12859-018-2367-z

67 Mu, J., Liu, H., Zhang, J., Luo, R. & Chen, H.-F. Recent force field strategies for intrinsically disordered proteins. *Journal of chemical information and modeling* **61**, 1037-1047 (2021). https://doi.org/https://doi.org/10.1021/acs.jcim.0c01175
68 Tzeng, G.-H. & Huang, J.-J. *Multiple attribute decision making: methods and applications*. (CRC press, 2011).
69 Saaty, T. L. *What is the analytic hierarchy process?* , (Springer, 1988).
70 Diakoulaki, D., Mavrotas, G. & Papayannakis, L. Determining objective weights in multiple criteria problems: The critic method. *Computers & Operations Research* **22**, 763-770 (1995). https://doi.org/https://doi.org/10.1016/0305-0548(94)00059-H
71 Mardani, A., Jusoh, A. & Zavadskas, E. K. Fuzzy multiple criteria decision-making techniques and applications – Two decades review from 1994 to 2014. *Expert Systems with Applications* **42**, 4126-4148 (2015). https://doi.org/https://doi.org/10.1016/j.eswa.2015.01.003
72 McKnight, P. E. & Najab, J. in *The Corsini Encyclopedia of Psychology* 1-1.
73 Xiang, S. *et al.* Phosphorylation Drives a Dynamic Switch in Serine/Arginine-Rich Proteins. *Structure* **21**, 2162-2174 (2013). https://doi.org/https://doi.org/10.1016/j.str.2013.09.014
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* **22**, 854-865 (2014). https://doi.org/https://doi.org/10.1016/j.str.2014.03.012
75 Chang, L., Mondal, A., Singh, B., Martínez-Noa, Y. & Perez, A. Revolutionizing peptide-based drug discovery: Advances in the post-AlphaFold era. *WIREs Computational Molecular Science* **14**, e1693 (2024). https://doi.org/https://doi.org/10.1002/wcms.1693
76 Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nature Communications* **14**, 2625 (2023). https://doi.org/10.1038/s41467-023-38328-5
77 Gainza, P. *et al.* De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176-184 (2023). https://doi.org/10.1038/s41586-023-05993-x
78 Khersonsky, O. & Fleishman, S. J. What Have We Learned from Design of Function in Large Proteins? *BioDesign Research* **2022**, 9787581 (2022). https://doi.org/doi:10.34133/2022/9787581
79 Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56-61 (2022). https://doi.org/doi:10.1126/science.add1964
80 and, D. S. G. & Olson, A. J. Structural Symmetry and Protein Function. *Annual Review of Biophysics and Biomolecular Structure* **29**, 105-153 (2000). https://doi.org/10.1146/annurev.biophys.29.1.105
81 An, L. *et al.* De novo design of diverse small molecule binders and sensors using Shape Complementary Pseudocycles. *bioRxiv*, 2023.2012.2020.572602 (2023). https://doi.org/10.1101/2023.12.20.572602
82 Rettie, S. A. *et al.* Cyclic peptide structure prediction and design using AlphaFold. *bioRxiv*, 2023.2002.2025.529956 (2023). https://doi.org/10.1101/2023.02.25.529956
83 Alamdari, S. *et al.* Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023.2009.2011.556673 (2023). https://doi.org/10.1101/2023.09.11.556673
84 Chu, A. E., Lu, T. & Huang, P.-S. Sparks of function by de novo protein design. *Nature*

Biotechnology **42**, 203-215 (2024). https://doi.org/10.1038/s41587-024-02133-2

85 Frank, C. *et al.* Efficient and scalable de novo protein design using a relaxed sequence space. *bioRxiv*, 2023.2002.2024.529906 (2023). https://doi.org/10.1101/2023.02.24.529906

86 Jeliazkov, J. R., Alamo, D. d. & Karpiak, J. D. ESMFold Hallucinates Native-Like Protein Sequences. *bioRxiv*, 2023.2005.2023.541774 (2023). https://doi.org/10.1101/2023.05.23.541774

87 Verkuil, R. *et al.* Language models generalize beyond natural proteins. *bioRxiv*, 2022.2012.2021.521521 (2022). https://doi.org/10.1101/2022.12.21.521521

88 Ji, X., Liu, H., Zhang, Y., Chen, J. & Chen, H.-F. Personal Precise Force Field for Intrinsically Disordered and Ordered Proteins Based on Deep Learning. *Journal of Chemical Information and Modeling* **63**, 362-374 (2023). https://doi.org/10.1021/acs.jcim.2c01501

89 Cui, X., Liu, H. & Chen, H.-F. Polarizable Force Field of Intrinsically Disordered Proteins with CMAP and Reweighting Optimization. *Journal of Chemical Information and Modeling* **62**, 4970-4982 (2022). https://doi.org/10.1021/acs.jcim.2c00835

90 Song, G., Zhong, B., Zhang, B., Rehman, A. U. & Chen, H.-F. Phosphorylation Modification Force Field FB18CMAP Improving Conformation Sampling of Phosphoproteins. *Journal of Chemical Information and Modeling* **63**, 1602-1614 (2023). https://doi.org/https://doi.org/10.1021/acs.jcim.3c00112

91 Zhu, J. *et al.* Phanto-IDP: compact model for precise intrinsically disordered protein backbone generation and enhanced sampling. *Brief Bioinform* **25** (2023). https://doi.org/10.1093/bib/bbad429

92 Haiyan, L., Yufeng, L. & Linghui, C. Diffusion in a quantized vector space generates non-idealized protein structures and predicts conformational distributions. *bioRxiv*, 2023.2011. 2018.567666 (2023).