PDFBENCH: A BENCHMARK FOR *De novo* PROTEIN DESIGN FROM FUNCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Function-guided protein design is a crucial task with significant applications in drug discovery and enzyme engineering. However, the field lacks a unified and comprehensive evaluation framework. Current models are assessed using inconsistent and limited subsets of metrics, which prevents fair comparison and a clear understanding of the relationships between different evaluation criteria. To address this gap, we introduce PDFBENCH, the first comprehensive benchmark for function-guided *de novo* protein design. Our benchmark systematically evaluates eight state-of-the-art models on 16 metrics across two key settings: description-guided design, for which we repurpose the Mol-Instructions dataset, originally lacking quantitative benchmarking, and keyword-guided design, for which we introduce a new test set, *SwissTest*, created with a strict datetime cutoff to ensure data integrity. By benchmarking across a wide array of metrics and analyzing their correlations, PDFBENCH enables more reliable model comparisons and provides key insights to guide future research. [1]

1 INTRODUCTION

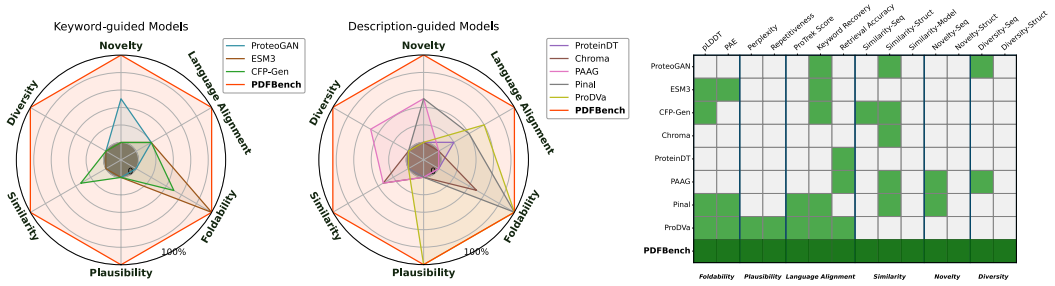
Proteins are essential macromolecules that play a key role in many biological processes by performing a wide range of functions. Protein design (Huang et al., 2016; Albanese et al., 2025) is of great significance in areas such as enzyme engineering (Planas-Iglesias et al., 2021; Kries et al., 2013) and drug discovery (Hung & Chen, 2014; Tiwari & Singh, 2022). Compared with unconditional generation (Ferruz et al., 2022), generation guided by specific functions (Lee et al., 2024) or control tags (Hayes et al., 2025), protein design based on user-specified functions provides greater practical value (Liu et al., 2025b; Madani et al., 2023). Current protein design from function tasks can be categorized into two types, description-guided (which utilizes textual functional descriptions as input) and keyword-guided (which employs function keywords such as InterPro (IPR) (Hunter et al., 2009) entries or Gene Ontology (GO) (gen, 2021) terms as input).

To build protein design models, choosing proper evaluation metrics is crucial. Various metrics (e.g., language alignment, foldability) are applied in the literature, but different models are usually evaluated in different ways. In Figure 1, we list existing evaluation metrics and depict how typical protein design models are evaluated against them. It shows that among the eight representative models, only half of them assess foldability, and only quarter of them evaluate novelty and diversity. For language alignment metrics, models either ignore model-based or retrieval-based metrics. Moreover, none of them consider all dimensions across the different metrics.

The lack of a comprehensive evaluation across all metrics can lead to unfair comparisons between methods. Furthermore, the correlations among different metrics have not been thoroughly investigated, limiting the understanding of these metrics and thereby hindering effective and insightful future research.

To bridge these gaps, we propose PDFBENCH, which, to the best of our knowledge, is the first benchmark designed to evaluate the capabilities of 8 novel function-guided *de novo* protein design methods, encompassing both description-guided approaches and keyword-guided approaches. Comprehensive benchmarking of the description-guided task is conducted using Mol-Instructions (Fang et al., 2023)

[1] The codes and datasets are available in the Anonymous GitHub repository: `https://anonymous.4open.science/r/PDFBench` and will be made publicly accessible in the final version.



(a) (b)

Figure 1: Overview of current function-guided protein design models evaluated using different metrics, highlighting the lack of a unified and comprehensive evaluation framework. (a) Proportion of metrics employed in each previous work. In PDFBENCH, metrics are categorized into 6 dimensions, and we show that none of the prior works have been evaluated across all dimensions. (b) Detailed view of the metrics in PDFBENCH, with several representative metrics from each dimension presented.

(*MolinstTest*), an instruction-following dataset of high quality but lacking any quantitative analysis of the designed proteins. For the keyword-guided task, we introduce a new test set, *SwissTest*, with a datetime cutoff to prevent data contamination. Furthermore, an in-depth analysis is conducted to explore the correlations among various evaluation metrics used in PDFBENCH.

In summary, the contributions of PDFBENCH are as follows:

- We present PDFBENCH, the first comprehensive benchmark for function-guided *de novo* protein design, encompassing both description-guided and keyword-guided settings (with 3 fine-grained tasks). We benchmark 8 state-of-the-art models across 16 metrics spanning 6 dimensions (i.e., Plausibility, Foldability, Language Alignment, Similarity, Novelty, and Diversity).
- We analyze and ensure fairness of evaluations. For the description-guided task, we repurpose Mol-Instructions in our benchmark (*MolinstTest*). We carefully analyze the testing configuration (by partitioning datasets based on potential overlaps) and demonstrate that fairness of *MolinstTest* results are reliable. For keyword-guided tasks, we introduce *SwissTest* which applies a strict datetime cutoff (e.g., only including SwissProt annotations after 2025) to ensure fairness.
- We identify key correlations between different metrics that facilitate fairer comparisons and provide insights for future model development. For instance, low PPL and Repeat scores consistently indicate well-folded protein structures (e.g., PPL exhibits Pearson correlations of 0.76 with pLDDT and -0.87 with PAE). Moreover, retrieval-based evaluations are highly sensitive to the chosen retrieval strategy, and although random sampling can be beneficial, absolute values should be interpreted with caution (e.g., the gap in Retrieval Accuracy for natural proteins across different retrieval strategies can reach 66.31%.).

2 PDFBENCH

2.1 TASKS AND DATASETS

As demonstrated in Figure 2, *de novo* protein design from function can be categorized into two types, description-guided and keyword-guided. The objective of both tasks is to generate novel proteins with specific functions, while the input format differs.

2.1.1 DESCRIPTION-GUIDED PROTEIN DESIGN

Task Definition The description-guided task is to design novel protein P with function description t written in natural language. The function description is to describe the overall function of P. The objective of this task is to generate a novel protein using the 20 standard amino acids $A = \{a_1, a_2, \cdots, a_{20}\}$.

$$p(P \mid t) = p((x_1, x_2, \cdots, x_k) \mid t, \forall i, x_i \in A)$$

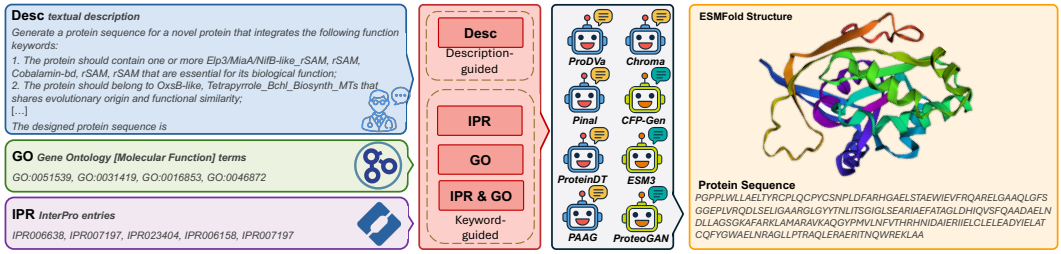


Figure 2: Examples of inputs and outputs for the description-guided protein design task and the keyword-guided protein design task (using GO and/or IPR keywords as inputs). Note that the GO and IPR terms can be converted into textual descriptions. A detailed explanation of this paradigm is provided in Appendix C.2.

Dataset Mol-Instructions (Fang et al., 2023) is a diverse, high-quality, large-scale instruction dataset for the biomolecular domain. We use its protein-oriented instruction test set as the evaluation set for the description-guided task, referred to as *MolinstTest*. The detailed construction process of *MolinstTest* is provided in Appendix C.1.

2.1.2 Keyword-guided Protein Design

Task Definition The keyword-guided task is to design a novel protein P based on a set of keywords $K = \{k_1, k_2, \cdots, k_n\}$, where each keyword k_i corresponds to either a Gene Ontology (molecular function) term or an InterPro term in PDFBench. Additionally, the keyword names are converted into textual function descriptions for comparison with models that accept text as input. The objective is to generate a novel protein based on K.

$$p(P \mid K) = p((x_1, x_2, \cdots, x_k) \mid K, \forall i, x_i \in A)$$

Dataset In contrast to the description-guided task, the input for the keyword-guided task consists of a series of keywords. In PDFBench, we manually curated a novel test dataset from UniProt/SwissProt UniProt Consortium (2018), termed *SwissTest*. To prevent potential data leakage, the inclusion period for *SwissTest* is restricted to January 1, 2025, through August 25, 2025. The detailed construction of *SwissTest* is presented in Appendix C.2.

2.1.3 Analysis for Fairness Evaluation in Benchmarking

It is important to ensure that the benchmark evaluation remains fair under specific controlled settings. To assess the impact of potential data contamination (i.e., overlaps [2] in protein functions or sequences within the training sets of the baseline models), we provide a detailed quantitative analysis in Appendix C.3. We ensure that PDFBench contains no hard overlaps by carefully selecting the test set (the Mol-Instructions test set is excluded from all baseline models) or by curating it (in *SwissTest*, we apply a datetime cutoff). For soft overlaps, our primary concern is that similar functions may potentially overlap. In Appendix C.3, we show that only 0.24% of the proteins in the Mol-Instructions test set share similar functions with at least 50% functional identity in SwissProt, which is primarily used by the baselines as the training set or its subsets. Further analysis demonstrates that excluding these soft overlaps in similar functions has minimal, or even positive, impact on the experimental results. This indicates that soft overlaps do not compromise the fairness of our evaluation.

2.2 Baselines

We ensemble recent *de novo* protein design from function baselines as shown in Table 1. For description-guided task, ProDVa (Liu et al., 2025a), Pinal (Dai et al., 2024), PAAG (Yuan et al., 2024), Chroma (Ingraham et al., 2023) and ProteinDT (Liu et al., 2025b) are supported. For keyword-guided

[2]We define hard overlaps as instances where an identical function–sequence pair appears in both the training and test sets. Additionally, soft overlaps refer to instances where the functions or sequences are similar.

task, ESM3 (Hayes et al., 2025), CFP-Gen (Yin et al., 2025) and ProteoGAN (Kucera et al., 2022) are supported. More details about these baseline are listed in Appendix E.

Table 1: Comparison with various baselines. The symbols ✓ and ✗ denote "Supported" and "Not Supported," respectively. As ProteoGAN, ESM3, and CFP-Gen offer only partial support for IPR and GO keywords, we report both the number of keywords each model claims to support and the number of keywords available in the *SwissTest* benchmark.

Baselines	Keyword			Desc	Training Set		Brief Summaries
	GO	IPR	GO&IPR		Size	Access	
ESM3	✗	✓ 29026 (1197)	✗	✗	455M	✗	A frontier multimodal generative model tokenizing sequence, structure, and function in unified space, trained with masked language modeling across modalities.
ProteoGAN	✓ 50 (41)	✗	✗	✗	158K	✗	Conditional GAN guided by Gene Ontology labels, designed for general-purpose protein generation and evaluated with biologically/statistically inspired metrics.
CFP-Gen	✓ 375 (92)	✓ 1154 (170)	✓	✗	244K	✓	Diffusion-based multimodal generator introducing AGFM and RCFE modules, enabling precise functional, structural, and residue-level control for multifunctional protein design.
ProteinDT	✓	✓	✓	✓	541K	✓	Multimodal framework using ProteinCLAP for joint embeddings, mapping text to protein representations, followed by autoregressive decoding for sequence generation.
Chroma	✓	✓	✓	✓	45K	✗	Diffusion-based protein designer integrating polymer physics, enabling programmable design through composable conditioners to enforce multiple constraints.
PAAG	✓	✓	✓	✓	129K	✓	A multimodal design model that employs a multilevel alignment module to align sequences and descriptions at both global and local levels.
Pinal	✓	✓	✓	✓	1.7B	✗	A large-scale framework, pretrained on 1.7B function–sequence pairs, most of which are synthetic, designing sequence mediated by structure.
ProDVa	✓	✓	✓	✓	640K	✓	A multimodal framework integrating a text encoder, a protein language model, and a fragment encoder that dynamically retrieves relevant fragments based on the specified function.

2.3 METRICS

Plausibility To evaluate the plausibility of the designed proteins, we introduce six metrics. Direct experimental validation of sequence plausibility is impractical for large-scale studies. Instead, we employ three protein language models—ProtGPT2 (Ferruz et al., 2022), ProGen2 (Nijkamp et al., 2023), and RITA (Hesslow et al., 2022)—to compute sequence perplexity. These yield ***PPL-ProtGPT2***, ***PPL-ProGen2***, and ***PPL-RITA***, which provide model-based estimates of how well the designed proteins conform to the distributional properties of natural proteins. Moreover, natural proteins rarely contain long repetitive fragments, yet prior studies (Ferruz et al., 2022; Wang et al., 2024b) indicate that generative models may produce sequences with abnormally high repetition, potentially impairing protein functionality. To assess this, we compute ***Rep-2*** and ***Rep-5*** following Rep-N (Welleck et al., 2019). In addition, we propose a new metric, ***Repeat***, which evaluates the fraction of repetitive sequence fragments from a biologically informed perspective.

Foldability We use ESMFold (Lin et al., 2022) to predict 3D structure for the designed sequence and present two metrics for accessing its foldability. Reasonable proteins must exhibit sufficient foldability to perform their functions. We compute the average predicted local distance difference test (***pLDDT***) and predicted aligned error (***PAE***) across the entire structure (Jumper et al., 2021; Akdel et al., 2022; Varadi et al., 2022).

Language Alignment Function-guided proteins should faithfully reflect user-specified properties. To evaluate the degree of alignment between textual descriptions and designed proteins, we employ five metrics: four model-based metrics (***ProTrek Score***, ***EvoLlama Score***, ***IPR Recovery*** and ***GO Recovery***), and one retrieval-based metric (***Retrieval Accuracy***). For model-based evaluation, we use ProTrek (Su et al., 2024), a multimodal protein language model pre-trained on large-scale protein–function pairs, to compute the ***ProTrek Score***, defined as the cosine similarity between protein embeddings and function embeddings. In addition, EvoLlama (Liu et al., 2024), fine-tuned on our dataset, is employed to derive the ***EvoLlama Score***, which measures the cosine similarity between the ground-truth function and the function predicted by EvoLlama. we calculate keyword-level recovery rates of designed proteins relative to natural proteins, specifically ***IPR Recovery*** and ***GO Recovery***. The IPR annotations are obtained using InterProScan (Paysan-Lafosse et al., 2023), while GO terms are derived from DeepGO-SE (Kulmanov et al.). Higher recovery rates indicate stronger preservation of functional characteristics. For retrieval-based evaluation, we follow previous work (Liu et al., 2025b) to compare each designed sequence against its ground-truth function and $T-1$ randomly selected functions using ProTrek. ***Retrieval Accuracy*** is defined as the proportion of cases in which the true function–sequence pair ranks highest among all comparisons.

Similarity Proteins with analogous sequences or structures often perform similar functions. ***ESM-Score*** is a composite metric comprising ESM-F1, ESM-Precision, and ESM-Recall, which measures sequence similarity between designed sequences and ground truth using ESM-2-650M (Lin et al., 2022), following the formulation of BERTScore (Zhang* et al., 2020). ***GT-Identity*** denotes the sequence similarity between a designed sequence and the ground truth, computed with MMseqs2 (Kallenborn et al., 2024), which employs multiple sequence alignment (MSA) to quantify similarity. Additionally, we assess structural similarity between designed proteins and ground truth using TMscore (Zhang & Skolnick, 2004), denoted as ***GT-TMscore***.

Novelty **Novelty** represents the dissimilarity of designed proteins with respect to reference databases. Here, we assess novelty from both sequence and structure perspectives. We employ MMseqs2 to calculate the novelty of designed sequences within UniProtKB (UniProt Consortium, 2018), denoted as ***Novelty-Seq***. Additionally, we utilize Foldseek (Kim et al.) to evaluate the novelty of designed protein structures within AlphaFoldDB/SwissProt, denoted as ***Novelty-Struct***.

Diversity Diversity evaluates whether the model generates diverse proteins rather than producing minor variations of a template sequence, by computing the average pairwise dissimilarity among all proteins designed for the same function. Similar to Novelty, we use MMseqs2 to compute sequence diversity (***Diversity-Seq***) and Foldseek to compute structural diversity (***Diversity-Struct***).

3 RESULTS

In this section, we conduct a comprehensive evaluation of these baselines across the two tasks (four fine-grained tasks). The detailed experimental settings are shown in Section F.1.

Table 2: Benchmark results for the description-guided task (Best, Second Best, Third Best). - indicates not applicable.

Models	Plausibility		Foldability		Language Alignment %			Novelty %		Diversity %	
	Perplexity ↓	Repeat % ↓	pLDDT ↑	PAE ↓	ProTrek Score ↑	EvoLlama Score ↑	Retrieval Accuracy ↑	Seq ↑	Struct ↑	Seq ↑	Struct ↑
Natural	5.99	1.99	80.64	9.20	27.00	60.33	89.01	4.90	13.56	-	-
Random(U)	21.71	0.72	22.96	24.85	1.03	36.22	13.03	58.14	77.64	97.01	81.59
Random(E)	18.68	1.15	25.77	24.71	1.04	34.11	12.73	60.19	76.82	99.56	81.45
ProteinDT	12.41	6.83	38.29	25.13	1.20	40.57	16.92	70.74	71.16	99.23	83.67
Chroma	12.19	2.59	59.18	15.03	2.10	40.10	13.01	58.68	51.06	96.13	79.90
PAAG	17.84	2.34	28.39	25.38	1.29	34.39	13.43	63.64	77.34	99.15	82.16
Pinal	5.81	12.83	75.25	10.96	17.50	53.40	63.43	43.82	17.23	82.96	72.73
ProDVa	7.63	1.92	76.86	8.66	17.40	51.19	66.83	14.64	36.31	83.29	36.92

3.1 DESCRIPTION-GUIDED

In Table 2, we report the benchmark results for the description-guided task on 11 main metrics. The complete results are displayed in Appendix F.2. Findings are summarized as follows:

(1) ProDVa can design relatively plausible sequence. Good sequence plausibility is fundamental to foldability and language alignment. The sequences designed by ProDVa exhibit repeat scores exceeding those of natural proteins and suboptimal perplexity scores in sequence rationality, indicating that ProDVa's design modules are capable of generating reasonable protein sequences.

(2) ProDVa and Pinal Generate Foldable Proteins. Both ProDVa and Pinal achieve substantially higher foldability scores compared with all other models. Specifically, ProDVa reaches the best pLDDT (76.86) and lowest PAE (8.66), while Pinal follows closely with pLDDT (75.25) and PAE (10.96). These results suggest that the sequences produced by both models are structurally stable and more likely to fold into valid conformations, highlighting the effectiveness of their design modules in capturing the structural constraints of proteins.

(3) ProDVa and Pinal exhibit comparable performance in Language Alignment, whereas the remaining baselines demonstrate substantially inferior results. For language alignment, both models outperform all baselines by large margins. ProDVa achieves the best retrieval accuracy

(66.83), while Pinal attains a comparable score (63.43). They also perform significantly better in ProTrek Score and EvoLlama Score compared to ProteinDT, Chroma, and PAAG. This indicates that the semantic and evolutionary information embedded in the descriptions are effectively translated into protein sequences by ProDVa and Pinal, whereas the baseline methods fail to capture such alignment.

(4) ProDVa and Pinal perform poorly with respect to novelty and diversity. While excelling in plausibility, foldability, and alignment, both models show relatively low novelty and diversity compared with baselines. The low novelty scores (ProDVa Seq/Struct: 14.64/36.31, Pinal Seq/Struct: 43.82/17.23) suggest that the designed sequences tend to remain close to the natural protein landscape. At the same time, their low diversity scores (ProDVa Seq/Struct: 83.29/36.92; Pinal Seq/Struct: 82.96/72.73) indicate that the models may confine functional design to narrow clusters in sequence/structure space. This reflects a trade-off: in order to achieve better functional alignment, ProDVa and Pinal may sacrifice exploration of diverse solutions, thereby limiting their coverage of the broader protein landscape.

Table 3: Benchmark results for the keyword-guided task (Best, Second Best, Third Best). - indicates not applicable.

Models	Plausibility		Foldability		Language Alignment %				Novelty %		Diversity %	
	Perplexity ↓	Repeat % ↓	pLDDT ↑	PAE ↓	ProTrek Score ↑	Retrieval Accuracy ↑	IPR Recovery↑	GO Recovery↑	Seq ↑	Struct ↑	Seq ↑	Struct ↑
guided with GO keywords												
Natural	9.17	2.17	76.92	10.54	21.60	77.49	100.00	77.49	4.07	18.15	-	-
Random(U)	21.74	0.72	23.20	24.56	4.29	0.00	20.79	10.00	58.04	76.75	94.29	81.56
Random(E)	18.68	1.14	25.99	24.47	3.44	0.00	11.71	11.06	60.28	75.92	98.65	81.54
ProteoGAN	18.03	2.50	28.72	24.67	4.42	0.00	14.99	13.84	65.24	75.82	98.94	84.37
CFP-Gen	5.16	12.67	73.38	14.61	10.03	9.67	18.98	38.87	47.85	28.28	85.14	81.76
ProteinDT	12.23	7.98	40.35	25.57	1.70	0.03	18.52	15.39	75.41	74.62	99.7	84.53
Chroma	12.18	2.71	59.27	15.00	1.84	0.23	16.33	12.07	59.35	50.88	93.70	79.79
PAAG	18.08	2.48	31.47	23.88	4.38	0.00	21.66	16.45	62.36	73.36	98.57	81.73
Pinal	6.85	14.13	72.58	11.79	12.69	19.26	22.76	49.93	46.06	19.27	87.61	79.00
ProDVa	11.16	1.87	74.73	6.11	14.42	20.22	30.24	52.38	25.02	32.72	98.17	35.76
guided with IPR keywords												
Natural	9.73	2.23	75.77	11.13	25.29	83.22	100.00	83.22	4.47	20.09	-	-
Random(U)	21.76	0.69	23.40	24.42	7.53	0.00	25.75	11.72	57.21	76.47	94.44	81.56
Random(E)	18.67	1.16	26.29	24.34	6.11	0.00	13.06	12.80	59.57	75.05	98.58	81.46
ESM3	6.33	28.13	60.90	16.73	6.22	20.17	15.43	33.01	71.87	37.56	91.41	76.79
CFP-Gen	4.94	11.86	76.36	12.54	10.21	32.79	23.41	40.96	49.46	23.15	85.31	82.08
ProteinDT	11.87	10.02	37.59	26.19	3.85	0.08	20.76	16.13	73.57	76.03	99.71	84.81
Chroma	12.17	2.60	59.76	14.67	3.82	0.17	17.15	13.68	59.25	50.77	94.06	79.88
PAAG	17.85	2.32	30.89	24.98	5.98	0.08	13.85	14.37	64.71	79.23	99.16	81.48
Pinal	8.12	16.73	65.69	14.10	14.38	25.63	15.93	57.59	51.61	27.00	87.02	80.38
ProDVa	12.47	1.99	72.80	6.86	15.19	24.58	26.59	51.99	28.86	31.26	95.02	45.60
guided with IPR&GO keywords												
Natural	8.96	2.16	77.17	10.48	27.36	87.39	100.00	87.39	3.89	17.72	-	-
Random(U)	21.72	0.73	22.85	24.72	4.84	0.00	25.38	9.45	57.48	77.18	94.24	81.73
Random(E)	18.68	1.14	25.60	24.59	3.72	0.00	14.67	11.28	60.06	75.98	98.39	81.57
CFP-Gen	5.23	13.14	72.70	14.45	11.68	35.21	23.31	45.77	54.72	28.89	80.61	81.91
ProteinDT	12.81	6.81	36.46	25.75	3.06	0.36	15.92	19.29	71.44	75.73	99.39	84.18
Chroma	12.19	2.53	58.71	15.33	2.19	0.16	14.12	11.67	59.36	51.45	94.27	79.97
PAAG	17.80	2.32	30.05	25.69	4.66	0.02	9.77	11.82	65.07	81.53	99.22	81.51
Pinal	7.39	16.22	69.32	12.97	15.26	33.08	21.64	60.88	49.03	22.43	85.20	78.20
ProDVa	10.48	2.61	74.26	8.06	16.78	30.95	25.24	61.23	21.97	24.20	91.94	52.18

3.2 KEYWORD-GUIDED

As illustrated in Table 3, we report the benchmark results for keyword-guided task on 12 main metrics. The complete results of keyword-guided task on all metrics are in Appendix F.3. Based on these results, our key findings are as follows:

(1) CFP-Gen, Pinal and ESM3 show great performance in Perplexity while the Repeat show poorly. These models achieve the lowest perplexity scores (CFP-Gen: 4.94–5.23; Pinal: 6.85–8.12; ESM3: 6.33), indicating that their generated sequences exhibit strong rationality under the protein language model. However, they also show much higher Repeat (ranging from 11.86 to 28.13) compared with ProDVa or Chroma, suggesting that the improved plausibility comes at the cost of local redundancy in sequence design.

(2) CFP-Gen, ProDVa and Pinal can design foldable proteins. These models consistently achieve high pLDDT and low PAE across different evaluation settings. ProDVa stands out with the best

overall foldability (pLDDT: 72.80–74.73; PAE: 6.11–8.06), while CFP-Gen and Pinal also produce structures with good confidence (pLDDT around 69–76; PAE around 11–14). This indicates that their design strategies are particularly effective at generating sequences that fold into stable 3D structures.

(3) CFP-Gen shows great performance among the keyword-guided baselines, while weak performance among the description-guided baselines. Compared to other keyword-guided baselines such as ProteinDT, Chroma, or PAAG, CFP-Gen achieves significantly higher alignment with biological annotations (e.g., IPR Recovery up to 35.21 and GO Recovery up to 21.05). In contrast, its performance was still less competitive than Pinal and ProDVa.

(4) Baselines perform better on IPR-guided task than GO-guided task. In the single-keyword setting, models achieve higher recovery performance on the metric that matches the input type. Moreover, using IPR as input generally leads to stronger performance across Plausibility, Foldability, and most Language Alignment metrics, indicating that IPR annotations provide more precise constraints for protein design than GO terms.

(5) The IPR&GO-guided task imposes stricter constraints than the other two tasks, while less than description-guided task. When extending from single- to dual-keyword guidance, we observe a nuanced trade-off. IPR Recovery increases while GO Recovery decreases, suggesting that IPR contributes more strongly to functional alignment in this joint setting. Meanwhile, both ProTrek Score and Retrieval Accuracy are improved, while Plausibility and Foldability remain largely unchanged. These results imply that combining IPR and GO constraints reduces the design difficulty in terms of language alignment, enabling models to better converge on functionally consistent sequences without sacrificing structural quality. Nevertheless, the alignment scores are still higher than those obtained in the description-guided setting, showing that structured keyword guidance provides clearer signals for functional targeting, albeit at the cost of reduced novelty and diversity.

4 Rethinking the Evaluation Metrics

In this section, we provide a comprehensive analysis regarding the correlations among different evaluation metrics on PDFBench.

4.1 Does PPL Accurately Reflect pLDDT and PAE?

We begin by exploring to what extent sequence-level metrics reflect protein structures. Folding proteins into 3D structures using AlphaFold (Jumper et al., 2021) or ESMFold (Lin et al., 2022) is time-consuming and requires substantial computational resources, particularly for longer sequences. Previous studies (Hesslow et al., 2022; Ferruz et al., 2022) have observed a correlation between PPL and pLDDT scores. However, no empirical results or further analyses have been conducted to investigate the correlation.

Results are randomly sampled from natural proteins with low PPL scores, Chroma-designed proteins with medium PPL scores, and randomly generated proteins with high PPL scores. Figure 3 presents the distributions of PPL, pLDDT, and PAE. Proteins with high pLDDT values are predominantly clustered in the low PPL range, whereas those with low pLDDT values are concentrated in the high PPL range. For the proteins situated between these two clusters, a negative correlation is observed between PPL and pLDDT values. Specifically, lower PPL values are generally associated with higher pLDDT scores. A similar pattern is observed in the distribution of PPL and PAE. Therefore, we empirically categorize PPL values into three ranges, denoted as low PPL range (values above 500), medium PPL range (values between 500 and 2,000), and high PPL range (values above 2,000). Additionally, the Pearson correlation (Cohen et al., 2009) in Figure 3(c) highlights the relationships between PPL, pLDDT, and PAE.

Takeaway I. In the low PPL range, proteins are well-folded, exhibiting high pLDDT scores and low PAE values. In contrast, proteins in the high PPL range struggle to fold into plausible structures. Within the medium PPL range, proteins with higher PPL values tend to display lower pLDDT scores and higher PAE values.

4.2 Do Repetitive Patterns Lead to Lower Structural Plausibility?

Previous research (Wang et al., 2024b) has found that repetitive patterns occurring in amino acid sequences may result in low pLDDT scores, thereby leading to lower structural plausibility. We conduct an empirical analysis using the Repeat metric to measure the correlation between this pattern and foldability. Proteins designed by ESM3 and Pinal exhibit repetitive patterns, as indicated by their high scores on the Repeat metric. Figure 4 presents the distribution of Repeat scores and pLDDT and PAE values for proteins randomly sampled from natural sequences and those designed by ESM3 and Pinal. One observation is that when the Repeat score remains relatively low, there is no clear relationship between Repeat and foldability. In other words, a low Repeat score does not necessarily indicate that a protein is well-folded. However, when the Repeat score exceeds 10, higher Repeat values are associated with lower pLDDT scores and higher PAE values. Therefore, it is important to maintain repetitive patterns below a certain threshold (e.g., Repeat < 10) when designing well-folded proteins.

Takeaway II. High Repeat scores (typically above 10) in protein sequences are associated with lower structural plausibility as indicated by lower pLDDT and higher PAE values.

4.3 How Faithfully Do Designed Proteins Align with Functional Descriptions?

The most reliable strategy for evaluating the alignment between designed proteins and input textual descriptions is through wet-lab experiments. However, such experiments are time-consuming and costly. Therefore, employing computational methods to screen proteins involves a trade-off between efficiency and accuracy. To more effectively evaluate the functions of designed proteins, both oracle model-based and retrieval-based metrics have been proposed.

We first investigate whether the two oracle model-based language alignment metrics exhibit consistency in evaluating natural proteins. These two metrics differ in two key perspectives. First, the ProTrek Score measures similarity between ground truth and designed proteins directly based on their embeddings, whereas the EvoLlama Score assesses similarity through predicted functional descriptions. Second, ProTrek is an oracle model pre-trained on large-scale datasets without further fine-tuning on specific downstream tasks. In contrast, EvoLlama is trained from scratch on the downstream task, leading to a distinct intrinsic knowledge distribution between the two oracle models. Figure 6(a) illustrates the consistency between the ProTrek Score and the EvoLlama Score for natural proteins sampled from the validation and test sets of our description-guided task. The results show that most proteins are accurately predicted and tightly clustered in the upper-right corner, indicating strong agreement between the two metrics. Furthermore, following (Dai et al., 2024) and the definitions introduced in Section 2.3, we establish empirical score thresholds to identify well-aligned proteins. Specifically, proteins with a ProTrek Score above 15 and an EvoLlama Score above 50 are considered to faithfully match the input functional descriptions.

Takeaway III. The ProTrek Score and the EvoLlama Score are two oracle-based metrics that demonstrate high agreement in evaluating protein functions. Proteins with a ProTrek Score above 15 and an EvoLlama Score above 50 are considered well-aligned, indicating they faithfully match the input functional descriptions.

The above discussion has remained focused on the global level of protein function. However, attention must also be directed toward local sequence alignment within proteins, particularly minor mutations in functional sub-sequences (motifs). To assess whether ProTrek is sensitive to protein mutations, we randomly select 1,000 natural proteins from *MolinstTest* and introduce random mutations with specified probabilities. The results are illustrated in Figure 5.

Takeaway IV. The ProTrek Score assesses both global alignment between the overall proteins and the functions, as well as local alignment between motifs and functions.

Next, we discuss the language alignment metrics that do not rely on oracle models. The GT-TM Score measures the similarity between a designed protein and the ground truth structure. Since

protein structure determines function, it is generally assumed that structurally similar proteins exhibit similar functions. However, we argue whether proteins with similar functions can fold into dissimilar structures. In Figures 6(b) and (c), proteins designed by Pinal and ProDVa are sampled for illustration. The average score reported in Table 2 is used as the threshold to determine whether the GT-TM score is considered high. It can be observed that 96.73% of the Pinal-designed proteins with high similarity to the ground truth exhibit high ProTrek scores (above 15), while 75.89% achieve high EvoLlama scores (above 50). A similar conclusion can be drawn from the ProDVa-designed proteins, demonstrating that high structural similarity leads to similar functions. Furthermore, for proteins with lower structural similarity, no correlation between the two similarities is observed.

Takeaway V. A high GT-TM Score generally indicates functional similarity among structurally similar proteins. However, high structural similarity is not a prerequisite for designing well-aligned proteins. Therefore, relying solely on this metric provides limited insight into whether the designed proteins align well with their functional descriptions.

In addition to the aforementioned metrics, Retrieval Accuracy is a retrieval-based metric that measures whether the embeddings of the positive function-sequence pair are the most similar among all candidates. However, this metric is highly dependent on the retrieved sequences. To assess the extent to which the retrieval strategy influences the results, we further define Soft Retrieval Accuracy and Hard Retrieval Accuracy. The difference between them lies in whether the $T-1$ most or least relevant texts and their corresponding sequences are retrieved in relation to the positive pair. The relevance between textual descriptions is defined by the cosine similarity of their embeddings. In Figure 7, for proteins designed by ProDVa and Pinal, the gap between Hard Retrieval Accuracy and Soft Retrieval Accuracy is 60.64% and 55.82%, respectively. Even for natural proteins, the gap between the two metrics on the ground truth can be as high as 66.31%. Therefore, the results demonstrate that the retrieval strategy significantly impacts performance.

Takeaway VI. The retrieval strategy employed in the Retrieval Accuracy metric has a significant impact on the evaluation results. Randomly sampling the negative pairs can serve as a workaround, but caution should be exercised when interpreting the absolute values of the metric.

5 Related Work

De novo protein design refers to the process of creating novel proteins from scratch, as opposed to the modification of existing sequences or structures (Huang et al., 2016). Current approaches in this field can be broadly categorized into unconditional and conditional generation methods. While unconditional generation operates without constraints, conditional generation methods—which guide model output using specific conditions—offer greater practical utility. Within the conditional generation domain, PDFBench primarily focuses on *de novo* protein design from function. The detail of related work is presented in B.

6 Conclusion

The field of protein design has experienced growing interest in recent years, particularly in function-guided approaches. However, the lack of comprehensive and efficient evaluation benchmarks has hindered progress in this area. To address this gap, we introduce PDFBench, a benchmark designed to evaluate *de novo* protein design from function. PDFBench focuses on four tasks and incorporates 16 metrics to ensure a fair and comprehensive assessment. Additionally, we analyze the utility of these metrics and their interrelationships, offering deeper insights into *de novo* protein design and the alignment between function and protein.

Reproducibility Statement

To ensure the reproducibility of PDFBench, we provide detailed descriptions of dataset curation (Section 2.1 and Appendix C), baselines (Appendix E), and experimental settings (Appendix F.1). Codes and datasets are available at the Anonymous GitHub.

References

The gene ontology resource: enriching a gold mine. *Nucleic acids research*, 49(D1):D325–D334, 2021.

Mehmet Akdel, Douglas EV Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, et al. A structural biology community assessment of alphafold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, 2022.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Bob Strome, Alan Moses, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pp. 2023–09, 2023.

Katherine I Albanese, Sophie Barbe, Shunsuke Tagami, Derek N Woolfson, and Thomas Schiex. Computational protein design. *Nature Reviews Methods Primers*, 5(1):13, 2025.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pp. 1–4, 2009.

Fengyuan Dai, Shiyang You, Chentong Wang, Yuliang Fan, Jin Su, Chenchen Han, Xibin Zhou, Jianming Liu, Hui Qian, Shunzhi Wang, et al. Toward de novo protein design from natural language. *bioRxiv*, pp. 2024–08, 2024.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

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

Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Zander Harteveld, Joshua Southern, Michaël Defferrard, Andreas Loukas, Pierre Vandergheynst, Micheal Bronstein, and Bruno Correia. Deep sharpening of topological features for de novo protein design. In *ICLR2022 Machine Learning for Drug Discovery*, 2022. URL `https://openreview.net/forum?id=DwN81YIXGQP`.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Yutong Hu, Yang Tan, Andi Han, Lirong Zheng, Liang Hong, and Bingxin Zhou. Secondary structure-guided novel protein sequence generation with latent graph diffusion. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 31–41. IEEE, 2024.

Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, 2016.

Che-Lun Hung and Chi-Chun Chen. Computational approaches for drug discovery. *Drug development research*, 75(6):412–418, 2014.

Sarah Hunter, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, et al. Interpro: the integrative protein signature database. *Nucleic acids research*, 37(suppl_1):D211–D215, 2009.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Felix Kallenborn, Alejandro Chacon, Christian Hundt, Hassan Sirelkhatim, Kieran Didi, Sooyoung Cha, Christian Dallago, Milot Mirdita, Bertil Schmidt, and Martin Steinegger. Gpu-accelerated homology search with mmseqs2. *bioRxiv*, pp. 2024–11, 2024.

Woosub Kim, Milot Mirdita, Eli Levy Karin, Cameron L. M. Gilchrist, Hugo Schweke, Johannes Söding, Emmanuel D. Levy, and Martin Steinegger. Rapid and sensitive protein complex alignment with Foldseek-Multimer. 22(3):469–472. ISSN 1548-7105. doi: 10.1038/s41592-025-02593-7. URL https://doi.org/10.1038/s41592-025-02593-7.

Hajo Kries, Rebecca Blomberg, and Donald Hilvert. De novo enzymes by computational design. *Current opinion in chemical biology*, 17(2):221–228, 2013.

Tim Kucera, Matteo Togninalli, and Laetitia Meng-Papaxanthos. Conditional generative modeling for de novo protein design with hierarchical functions. *Bioinformatics*, 38(13):3454–3461, 05 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac353. URL https://doi.org/10.1093/bioinformatics/btac353.

Maxat Kulmanov, Francisco J. Guzmán-Vega, Paula Duek Roggli, Lydie Lane, Stefan T. Arold, and Robert Hoehndorf. Protein function prediction as approximate semantic entailment. 6(2):220–228. ISSN 2522-5839. doi: 10.1038/s42256-024-00795-w. URL https://doi.org/10.1038/s42256-024-00795-w.

Minji Lee, Luiz Felipe Vecchietti, Hyunkyu Jung, Hyun Joo Ro, Meeyoung Cha, and Ho Min Kim. Robust optimization in protein fitness landscapes using reinforcement learning in latent space. *arXiv preprint arXiv:2405.18986*, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Nuowei Liu, Changzhi Sun, Tao Ji, Junfeng Tian, Jianxin Tang, Yuanbin Wu, and Man Lan. Evollama: Enhancing llms' understanding of proteins via multimodal structure and sequence representations. *arXiv preprint arXiv:2412.11618*, 2024.

Nuowei Liu, Jiahao Kuang, Yanting Liu, Changzhi Sun, Tao Ji, Yuanbin Wu, and Man Lan. Protein design with dynamic protein vocabulary, 2025a. URL https://arxiv.org/abs/2505.18966.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, et al. A text-guided protein design framework. *Nature Machine Intelligence*, pp. 1–12, 2025b.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8): 1099–1106, 2023.

Weian Mao, Muzhi Zhu, Zheng Sun, Shuaike Shen, Lin Yuanbo Wu, Hao Chen, and Chunhua Shen. De novo protein design using geometric vector field networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9UIGyJJpay.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic acids research*, 51(D1):D418–D427, 2023.

Joan Planas-Iglesias, Sérgio M Marques, Gaspar P Pinto, Milos Musil, Jan Stourac, Jiri Damborsky, and David Bednar. Computational design of enzymes for biotechnological applications. *Biotechnology Advances*, 47:107696, 2021.

Jin Su, Xibin Zhou, Xuting Zhang, and Fajie Yuan. Protrek: Navigating the protein universe through tri-modal contrastive learning. *bioRxiv*, pp. 2024–05, 2024.

Anshul Tiwari and Sakshi Singh. Computational approaches in drug designing. In *Bioinformatics*, pp. 207–217. Elsevier, 2022.

The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46 (5):2699–2699, 2018.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Chentong Wang, Yannan Qu, Zhangzhi Peng, Yukai Wang, Hongli Zhu, Dachuan Chen, and Longxing Cao. Proteus: exploring protein structure generation for enhanced designability and efficiency. *bioRxiv*, pp. 2024–02, 2024a.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *International Conference on Machine Learning*, 2024b.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training, 2019. URL https://arxiv.org/abs/1908.04319.

Junbo Yin, Chao Zha, Wenjia He, Chencheng Xu, and Xin Gao. CFP-gen: Combinatorial functional protein generation via diffusion language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EiM163eZyg.

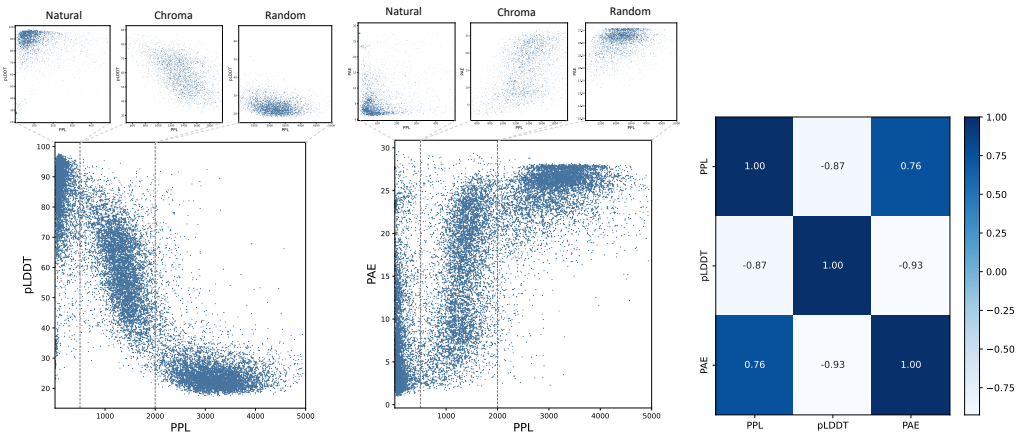
Chaohao Yuan, Songyou Li, Geyan Ye, Yikun Zhang, Long-Kai Huang, Wenbing Huang, Wei Liu, Jianhua Yao, and Yu Rong. Annotation-guided protein design with multi-level domain alignment. *arXiv preprint arXiv:2404.16866*, 2024.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

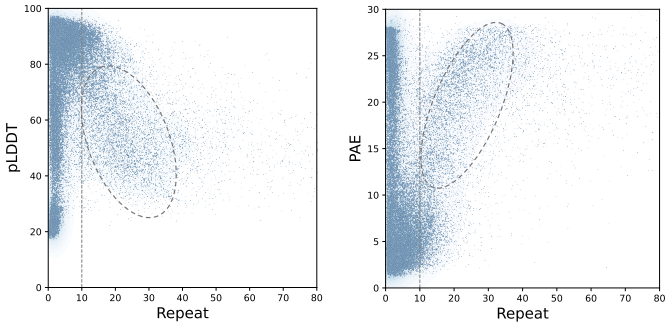
Yinuo Zhang, Phil He, Ashley Hsu, Sophia Vincoff, and Pranam Chatterjee. Ppi-llama2: De novo generation of binding proteins conditioned on target sequence alone. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*.

A Additional Figures



(a) (b) (c)

Figure 3: **(a)** presents the distribution of PPL and pLDDT. **(b)** displays the distribution of PPL and PAE. **(c)** illustrates the Pearson correlation among these metrics. Note that PPL values are categorized into three ranges: values below 500 indicate a low PPL range, values between 500 and 2,000 represent a medium PPL range, and values above 2,000 correspond to a high PPL range.



(a) (b)

Figure 4: **(a)** Distribution of Repeat and pLDDT. **(b)** Distribution of Repeat and PAE.

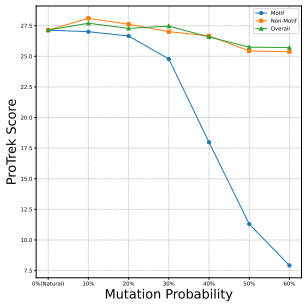
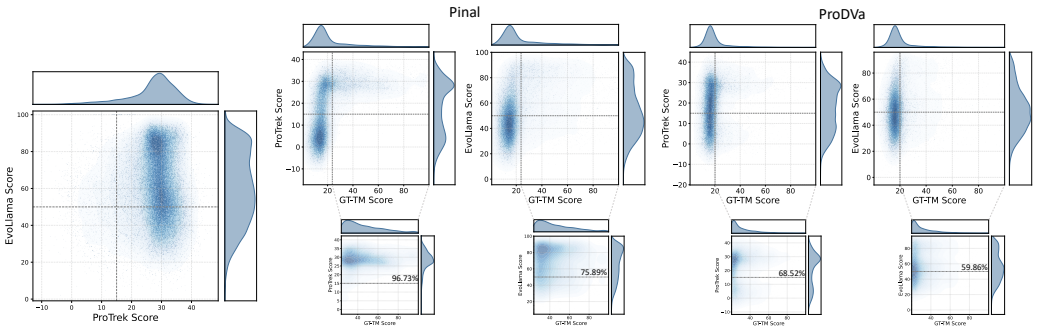


Figure 5: Results for random mutations in natural proteins. Motif and Non-motif indicate mutations within or outside motif regions. Overall includes all mutations.

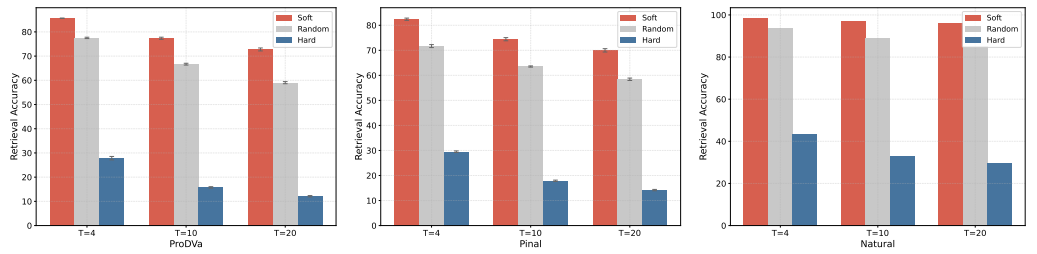
B *De novo* Protein Design

De novo protein design aims to generate novel proteins and can be categorized into unconditional and conditional approaches.



(a) (b) (c)

Figure 6: **(a)** presents the distribution of the ProTrek Score and EvoLlama Score for natural proteins. **(b)** and **(c)** present the distributions of the GT-TM Score, ProTrek Score, and EvoLlama Score for proteins designed by Pinal and the ProDVa.



(a) (b) (c)

Figure 7: Experimental results are reported for Soft, Random, and Hard Retrieval Accuracy. The random variant refers to the original implementation.

Unconditional *de novo* Unconditional design methods generate amino acid sequences without constraints, employing either autoregressive models (Madani et al., 2023; Ferruz et al., 2022; Hesslow et al., 2022) or discrete diffusion models (Alamdari et al., 2023; Wang et al., 2024b). Alternatively, some methods (Watson et al., 2023; Mao et al., 2024; Wang et al., 2024a) adopt a two-stage paradigm: first generating backbone structures through diffusion in SE(3) space, then predicting corresponding sequences. Additionally, certain approaches (Ren et al., 2024; Wang et al., 2024) utilize diffusion processes or energy-based models to jointly generate both backbone structures and sequences.

Conditional *de novo* Conditional *de novo* design methods incorporate additional information to guide the design process, such as specified three-dimensional structures (Harteveld et al., 2022; Dauparas et al., 2022; Hsu et al., 2022; Gao et al., 2022), secondary structures (Hu et al., 2024), protein-protein interactions (Zhang et al.), control tags (Nijkamp et al., 2023), or function.

***De novo* protein design from function** *De novo* protein design from function has recently garnered substantial interest, Existing methods can be categorized as description-guided methods and keyword-guided methods. Description-guided methods include Chroma (Ingraham et al., 2023), ProteinDT (Liu et al., 2025b), PAAG (Yuan et al., 2024), Pinal (Dai et al., 2024), ProDVa (Liu et al., 2025a). Chroma utilizes the diffusion framework to simultaneously design protein structure and sequence. ProteinDT employs a contrastive learning paradigm to align the representation spaces of function and protein, achieving over 90% retrieval accuracy on its own evaluation. PAAG introduces a multi-level alignment module that enables simultaneous attention to both protein-level and domain-level information, resulting in superior performance on success rate metrics compared to previous approaches. Pinal differs from the end-to-end generation approaches mentioned above by first generating the protein structure based on function, followed by sequence design informed by both function and structure. Notably, Pinal utilizes the largest dataset, containing over 1000 times more protein-function pairs than all other methods. ProDVa employs a novel mechanism

to dynamically retrieve the most relevant fragments from natural protein sequences based on the target function, significantly enhancing the method's performance. Keyword-guided methods include ProteoGAN (Kucera et al., 2022), ESM3 (Hayes et al., 2025) and CFP-Gen (Yin et al., 2025). ProteoGAN employ a generative adversarial networks (GAN) to design protein sequences for 50 Gene Ontology terms. ESM3, a cutting-edge multi-modal generative language model, integrates reasoning over protein sequence, structure, and function. It can respond to complex prompts combining these modalities and demonstrates high alignment responsiveness, enhancing its predictive fidelity. For keyword-guided task, it can generate sequences for 29026 IPR entries (nearly three-fifths of the InterPro). CFP-Gen utilizes the latest generative framework-masked diffusion model-to design protein sequences for 375 GO terms and 1154 IPR entries. As has been previously mentioned, all of the aforementioned methods claim to achieve optimal performance in their own reviews (i.e. their own proposed metrics). The objective of PDFBENCH is to provide a fair and comprehensive benchmark for evaluating the performance of the 8 novel methods.

C DETAILS OF DATASETS

C.1 *MolinstTest*

Mol-Instructions (Fang et al., 2023) comprises three key categories of instructions: molecule-oriented, protein-oriented, and biomolecular text. The protein-oriented instructions[3] include a 196K subset, Protein Design, for *de novo* protein design, sourced from UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. We use the test set of this subset with 5,876 proteins for our evaluation, referred to as *MolinstTest*.

C.2 *SwissTest*

Unlike description-guided tasks, keyword-guided tasks lack a publicly available, high-quality evaluation dataset. To address this gap, we constructed a novel dataset. Specifically, we selected proteins from UniProt/SwissProt[4] released between January 1, 2025, and August 25, 2025, and collected their corresponding protein sequences, InterPro IDs, and Gene Ontology terms[5]. To allow description-guided baselines to participate in the evaluation, we followed the approach of Mol-Instructions (Fang et al., 2023) by concatenating the text descriptions associated with InterPro entries and Gene Ontology terms using a prompt. The detailed processing steps are provided in Table 4. We finally curated a novel dataset containing 1,057 proteins with 1,297 IPR entries and 380 GO terms, referred to as *SwissTest*.

Table 4: Prompt and templates for converting keywords into textual description.

Prompt	
Generate a protein sequence for a novel protein that integrates the following function keywords: {textual annotations for all keywords connected by semicolons}. The designed protein sequence is	
Keyword Type	**Template**
InterPro (Domain)	The protein should contain one or more {`domains`} that are essential for its biological function
InterPro (Family)	The protein should belong to {`family`*s*} that shares evolutionary origin and functional similarity
InterPro (Homologous_Superfamily)	The protein should be classified within {`homologous superfamily`*s*} sharing conserved structural features
InterPro (Repeat)	The protein should include one or more {`repeat`} that provide structural or functional support
InterPro (Conserved_Site)	The protein should contain {`conversed site`*s*} that is preserved across related proteins
InterPro (Active_Site)	The protein must have {`activate sites`} that is conserved among related catalytic enzymes
InterPro (Binding_Site)	The protein should include a {`binding site`*s*} that enables ligand binding under diverse conditions
InterPro (PTM)	The protein should contain {`PTM(s)`} that allow regulation through chemical modifications
Gene Ontology (Molecular Function)	The protein must be able to perform the {molecular function*s*} required for its activity

[3] `https://huggingface.co/datasets/zjunlp/Mol-Instructions`

[4] `https://www.uniprot.org/uniprotkb?query=*`

[5] Search with `(date_created:[2025-01-01 TO 2025-08-25]) AND (reviewed:true)`

C.3 Fairness Analysis

For the description-guided task, the *MolinstTest* is excluded from the training process of all selected baseline models, including ProteinDT (Liu et al., 2025b), Chroma (Ingraham et al., 2023), PAAG (Yuan et al., 2024), Pinal (Dai et al., 2024), and ProDVa (Liu et al., 2025a). For the keyword-guided task, we apply a datetime cutoff and exclude all proteins and their corresponding keyword annotations dated prior to January 1, 2025, thereby strictly preventing potential data leakage. Therefore, there is no risk of data contamination (hard overlaps) during the evaluation.

While we also consider soft overlap settings (i.e., cases involving similar functions or sequences), in PDFBench, functions are provided as inputs, and the model is tasked with designing novel proteins (outputs) aligned with these functions. The model's predictions are conditioned solely on the provided functions, and any overlap in outputs across splits neither confers an unfair advantage nor compromises the integrity of the evaluation. Therefore, the presence of similar proteins does not constitute data contamination in the conventional sense and has minimal impact on the evaluation. For soft overlaps in similar functions, we conduct experiments to validate the fairness of our evaluation on the three baselines with publicly available training set, including ProDVa (Liu et al., 2025a), ProteinDT (Liu et al., 2025b), and PAAG (Yuan et al., 2024). First, we use ProTrek (Su et al., 2024) to compute the embeddings of the functions and sequences in the three training sets and *MolinstTest*. Then, for each function–sequence pair in *MolinstTest*, if an identical sequence exists in the corresponding training set, we compute the cosine similarity between ProTrek embeddings of the function in *MolinstTest* and the corresponding function in the training set. If the similarity exceeds the `Threshold`, it is considered a soft overlap. We categorize the baseline results on *MolinstTest* into two groups based on the presence or absence of overlap, as shown in Table 5. The analysis shows that, for ProDVa and ProteinDT, excluding soft overlaps has a positive impact on Perplexity (up to 2.12%), ProTrek Score (up to 2.74%), and Retrieval Accuracy (up to 4.56%). We observe minimal impact on foldability, despite slight improvements in pLDDT and PAE when soft overlaps are included. Additionally, for PAAG, the impact of including soft overlaps is consistently minimal across all metrics.

Table 5: Soft overlap analysis on *MolinstTest*. Results including soft overlaps are presented in parentheses, while those excluding soft overlaps are shown outside the parentheses.

Models	#Threshold (#Num)	Perplexity ↓	pLDDT ↑	PAE ↓	ProTrek Score ↑	Retrieval Accuracy ↑
ProDVa	0.0 (5876)	- (7.63)	- (76.86)	- (8.66)	- (17.40)	- (66.83)
	0.1 (1626)	7.13 (8.95)	76.75 (77.06)	9.08 (7.61)	18.16 (15.42)	67.93 (63.37)
	0.3 (185)	7.56 (9.68)	76.81 (77.61)	8.72 (7.42)	17.49 (14.78)	66.83 (61.62)
	0.5 (14)	7.63 (9.63)	76.83 (79.21)	8.68 (7.51)	17.41 (14.71)	66.67 (66.67)
	0.7 (2)	7.63 (5.54)	76.84 (78.23)	8.68 (5.67)	17.40 (23.80)	66.66 (100.00)
ProteinDT	0.0 (5876)	- (12.41)	- (38.29)	- (25.13)	- (1.20)	- (16.92)
	0.1 (1524)	12.45 (12.29)	37.50 (40.55)	25.12 (25.17)	1.06 (1.59)	16.35 (17.15)
	0.3 (112)	12.40 (12.49)	38.26 (39.77)	25.13 (25.29)	1.20 (1.05)	16.60 (14.58)
	0.5 (5)	12.41 (11.55)	38.29 (44.35)	25.13 (26.21)	1.20 (1.96)	16.56 (13.33)
	0.7 (0)	-	-	-	-	-
PAAG	0.0 (5876)	- (17.84)	- (28.39)	- (25.38)	- (1.29)	- (13.43)
	0.1 (749)	17.85 (17.79)	28.37 (28.53)	25.36 (25.47)	1.18 (2.00)	12.65 (14.11)
	0.3 (698)	17.85 (17.79)	28.38 (28.49)	25.36 (25.51)	1.20 (1.89)	12.67 (14.04)
	0.5 (579)	17.85 (17.79)	28.40 (28.29)	25.36 (25.55)	1.24 (1.74)	12.67 (14.28)
	0.7 (318)	17.84 (17.83)	28.40 (28.20)	25.37 (25.57)	1.27 (1.53)	12.70 (15.09)

D Details of Metrics

D.1 Plausibility

Repetitiveness We use the ***RepN*** metric to reflect the repetition at n-gram levels in the designed protein sequence P. Additionally, we propose a metric, ***Repeat***, to more accurately evaluate the proportion of repetitive sequence fragments from a biological perspective.

$$\text{Rep-n} = 100 \times (1.0 - \frac{\text{unique n-grams}(P)}{\text{total n-grams}(P)})$$

Algorithm 1 Compute Repeat

Require: $sequence \neq \emptyset$
Ensure: $proportion \in [0.0, 1.0]$
  $n \leftarrow sequence$
  if $n = 0$ **then**
    return 0.0
  end if
  $regions \leftarrow \emptyset$
  $max\_window\_size \leftarrow \min(20, \lfloor n/2 \rfloor)$
  for $window\_size = 1$ **to** max_window_size **do**
    for $i = 0$ **to** $n - window\_size$ **do**
      $pattern \leftarrow sequence[i : i + window\_size]$
      $count \leftarrow 1$
      $j \leftarrow i + window\_size$
      while $j \leq n - window\_size$ **and** sequence[j:j+window_size] = pattern **do**
        $count \leftarrow count + 1$
        $j \leftarrow j + window\_size$
      end while
      if $count \geq 3$ **then**
        $regions \leftarrow regions \cup \{(i,\ i + window\_size \times count)\}$
      end if
    end for
  end for
  if $regions = \emptyset$ **then**
    return 0.0
  end if
  $merged \leftarrow$ sorted and merged regions
  $total\_repeat \leftarrow \sum_{(start,end) \in merged}(end - start)$
  $proportion \leftarrow total\_repeat/n$
  return $proportion$

D.2 Language Alignment

ProTrek Score Given a designed sequence P and respective description T, ***ProTrek score*** is defined as:

$$\text{ProTrek Score} = \cos(\tau_{\text{seq}}(P), \tau_{\text{text}}(T))/t$$

where$\tau_{\text{seq}}(P)$, $\tau_{\text{text}}(T)$ and t represent the protein sequence encoder, text encoder and temperature of ProtTrek, respectively.

EvoLlama Score In addition to the ProTrek Score, we employ a generative approach for alignment evaluation using EvoLlama (Liu et al., 2024). Specifically, we utilize EvoLlama, which comprises a 650M ESM2 protein sequence encoder and a 3B Llama-3.2 text decoder. The model is randomly initialized and trained from scratch using the SwissMolinst dataset described in Liu et al. (2025a). Given the designed protein P and respective function description t, we prompt the P with `The function of the protein is` and fetch EvoLlama-generated function description t' with it. Assume that the t and t' can be tokenized into k and k' tokens, respectively. The ***EvoLlama Score*** is defined as follows:

$$\text{EvoLlama Score} = \text{sim}(\frac{1}{k}\sum_{i=1}^{k}\texttt{Embed}(t), \frac{1}{k'}\sum_{i=1}^{k'}\texttt{Embed}(t'))$$

where $\texttt{Embed}(\cdot)$ denotes using PubMedBERT (Gu et al., 2021) as the embedding model.

Keyword Recovery Given the designed sequence P and ground truth sequence GT, ***IPR Recovery*** and ***GO Recovery*** are defined as:

$$\text{IPR Recovery} = \begin{cases} \dfrac{\text{InterProScan}(P) \cap \text{InterProScanE}(GT)}{\text{InterProScan}(GT)}, & \text{if InterProScanO}(GT) \neq \emptyset, \\ \text{N/A}, & \text{if InterProScan}(GT) = \emptyset. \end{cases}$$

$$\text{GO Recovery} = \begin{cases} \dfrac{\text{DeepGO-SE}(P) \cap \text{DeepGO-SE}(GT)}{\text{DeepGO-SE}(GT)}, & \text{if DeepGO-SE}(GT) \neq \emptyset, \\ \text{N/A}, & \text{if DeepGO-SE}(GT) = \emptyset. \end{cases}$$

Retrieval Accuracy ***Retrieval Accuracy*** is obtained with a well-pretrained model, i.e. ProTrek: Given the designed sequence P, respective description T_0 and N randomly selected descriptions as negative pool $\mathcal{N} = (T_1, T_2, ..., T_{N-1})$ from testing set, retrieval accuracy is defined as:

$$\text{Retrieval Accuracy} = [\cos(\tau_{\text{seq}}(P), \tau_{\text{text}}(T_0)) \geq \cos(\tau_{\text{seq}}(P), \tau_{\text{text}}(T_i)), \quad \forall i \in \mathcal{N}]$$

D.3 SIMILARITY

GT-Identity Given the designed sequence P and ground truth sequence GT, we compute the ***GT-Identity*** with the align module of MMseqs (Kallenborn et al., 2024).

$$\text{GT-Identity} = \text{MMseqs}^{\text{align}}(P, GT)$$

ESMScore Following the formula in (Zhang* et al., 2020), we compute the BertScore between the ground truth sequence GT and the designed sequence P using ESM-2-650M (Lin et al., 2022), namely ***ESMScore***.

GT-TMscore We measures the ***GT-TMscore*** between the ESMFold-predicted structures between the design sequence P and the ground truth sequence GT using TMscore.

$$\text{GT-TMscore} = \text{TMscore}(\text{Structure}_P, \text{Structure}_{GT})$$

D.4 NOVELTY

Sequence Novelty Initially, MMseqs2 is utilized to retrieve the num_prot most similar sequences and the respective similarity sim_i of the designed sequence P to UniProtKB. Subsequently, it is possible to obtain each novelty nov_i via $1 - sim_i$ if a match sequence is present, otherwise 1. Finally, ***Novelty-Seq$_{Easy}$*** and ***Novelty-Seq$_{Hard}$*** can be defined as:

$$\text{Novelty-Seq}_{\text{Easy}} = \frac{\sum nov_i}{num\_prot}, \ \text{Novelty-Seq}_{\text{Hard}} = 1 - \max_i sim_i$$

Stucture Novelty Similar to the Sequence Novelty, given the designed sequence P, Foldseek (Kim et al.) is used to retrieve the num_prot most similar structures and their respective similarities sim_i between the ESMFold-predicted structure $Structure_P$ to AlphafoldDB/SwissProt[6].The values of $nov_i$are then obtained in the same manner as for Sequence Novelty. Finally, ***Novelty-Struct$_{Easy}$*** and ***Novelty-Struct$_{Hard}$*** can be defined as:

$$\text{Novelty-Struct}_{\text{Easy}} = \frac{\sum nov_i}{num_{prot}}, \ \text{Novelty-Struct}_{\text{Hard}} = 1 - \max_i sim_i$$

[6] https://alphafold.ebi.ac.uk/download#swissprot-section

D.5 DIVERSITY

Sequence Diversity Given the N sequences $\mathcal{P} = (P_1, P_2, ..., P_N)$ designed from the same description, we employ MMseqs2 to compute the similarity between each pair of sequences in $\mathcal{P}$. The same as Novelty, we obtain each diversity div_i via $1 - sim_i$ if the similarity between two sequences can be computed, otherwise 1, which means the two sequences are totally different. Finally, ***Diversity-Seq*** is defined as:

$$\text{Diversity-Seq} = \frac{\sum div_i}{N(N-1)}$$

Structure Diversity Given the N sequences $\mathcal{P} = (P_1, P_2, ..., P_N)$ designed from the same description and their corresponding ESMFold-predicted structures $\mathcal{S} = (S_1, S_2, ..., S_N)$, we use Foldseek to compute pairwise structural similarities within $\mathcal{S}$, , yielding similarity scores sim_i and corresponding diversity values div_i. Finally, ***Diversity-Struct*** is defined as:

$$\text{Diversity-Struct} = \frac{\sum div_i}{N(N-1)}$$

E DETAILS OF BASELINES

E.1 DESCRIPTION-GUIDED DESIGN BASELINES

Chroma Chroma (Ingraham et al., 2023) is a diffusion-based and programmable generation model for proteins. It employs a structured diffusion process that takes the physical properties of proteins as polymer chains into consideration. Chroma not only generates protein sequences using functional keywords, but it is also highly "programmable": users can guide the generation process using composable conditioners to enforce constraints such as symmetry, substructure, shape, or semantic properties (e.g., protein class, text prompts).

ProteinDT ProteinDT (Liu et al., 2025b) is a multimodal framework that utilizes function descriptions to guide protein design. The fundamental principle underpinning the system is ProteinCLAP (Contrastive LAnguage and Protein), which employs comparative learning to align description and protein sequence representations within a shared embedding space. The framework is composed of three sequential steps: ProteinCLAP aligns the representation, a facilitator maps the text embedding to the protein representation, and a autoregressive decoder generates a sequence based on this representation.

PAAG PAAG (Yuan et al., 2024) is a multimodal framework that focuses on both functional descriptions and functional keywords. It employs a multilevel alignment module that aligns protein sequences and description/keywords at the global and local levels. Subsequently, an autoregressive decoder is employed to generate protein sequences based on the aligned annotation representations.

Pinal Pianl (Dai et al., 2024) is a large-scale(up to 16B parameters) *de novo* protein design framework intended to translate natural language instructions into novel protein sequences. In lieu of direct text-to-sequence generation, Pinal adopts a two-stage approach: first, protein structures are generated from linguistic descriptions; then, sequences are designed based on the generated structures and the original linguistic input. This strategy employs the relatively restricted structure space as a preliminary step to efficiently constrain the extensive sequence search space.

ProDVa ProDVa (Liu et al., 2025a) is a multimodal protein design framework that combines textual function descriptions with insights from natural protein fragments to create sequences that are both functionally aligned and structurally plausible. It integrates a text encoder, a protein language model, and a fragment encoder that dynamically retrieves the most relevant fragments based on the desired function.

E.2 KEYWORD-GUIDED DESIGN BASELINES

ProteoGAN ProteoGAN (Kucera et al., 2022) is a conditional generative adversarial network designed to generate novel protein sequences based on functional labels from the Gene Ontology (GO).

ESM3 ESM3 (Hayes et al., 2025) is a large-scale (up to 98 billion parameters) multimodal generative language model designed to simulate protein evolution. It represents protein sequence, structure, and function as discrete tokens processed within a unified latent space. ESM3 is trained using a generative masked language modeling (MLM) objective, predicting randomly masked tokens across modalities to learn their complex interrelationships. The model can follow complex prompts combining sequence, structure, and function information.

CFP-Gen CFP-Gen (Yin et al., 2025) is a novel diffusion language model designed for generating functional proteins by simultaneously integrating multiple constraints from different modalities, including function, sequence, and structure. It employs an Annotation-Guided Feature Modulation (AGFM) module to control protein features using functional annotations and a Residue-Controlled Functional Encoding (RCFE) module for precise residue-level control.

F DETAILS OF RESULTS

F.1 EXPERIMENT SETTING

We implement the baselines as follows,

Description-guided For description-guided baselines (ProDVa, Pinal, PAAG, Chroma, ProteinDT), we directly use the description as input to prompt the model to generate sequences.

Keyword-guided Keyword-guided baselines only support restricted keywords, necessitating additional data processing. For InterPro entries unsupported by ESM3 and CFP-Gen, we opted to skip these keywords to ensure comparative fairness. For Gene Ontology terms unsupported by ProteoGAN and CFP-Gen, we attempted to find their ancestor terms in the GO database for substitution; if none were found, we skipped the term.

All baselines We evaluated using the official baseline implementation and weights provided. The results are averaged across three runs with different random seeds to ensure fairness. Additionally, no further fine-tuning is performed on any of the baselines to ensure a fair evaluation of the models' capability to design novel and functional proteins under the same settings.

F.2 COMPLETE RESULTS FOR THE DESCRIPTION-GUIDED TASK

Results on keyword-guided task are in Table 6, Table 7 and Table 8. We mark the top three models, with deeper colors indicating superior performance.

Table 6: Benchmark results of Plausibility and foldability for the description-guided task.

Models	Perplexity			Repetitveness			Foldability			
	PPL-ProtGPT2 ↓	PPL-ProGen ↓	PPL-RITA ↓	Repeat ↓	Rep-2 ↓	Rep-5 ↓	pLDDT ↑	% > 70 ↑	PAE ↓	% < 10 ↑
Natural	318.15	5.99	5.52	1.99	44.49	0.25	80.64	81.16	9.20	65.64
Random(U)	2484.04±4.53	21.71±0.00	22.14±0.01	0.72±0.01	34.59±0.03	0.01±0.00	22.96±0.04	0.16±0.04	24.85±0.01	0.56±0.03
Random(E)	3136.88±4.17	18.68±0.00	19.04±0.00	1.15±0.01	40.99±0.01	0.01±0.00	25.77±0.03	0.19±0.06	24.71±0.01	0.60±0.03
ProteinDT	1576.23±4.32	12.41±0.01	12.44±0.01	6.83±0.10	62.47±0.14	2.82±0.05	38.29±0.04	0.98±0.17	25.13±0.02	0.40±0.09
Chroma	1370.21±1.48	12.19±0.00	12.42±0.01	2.59±0.02	55.41±0.03	0.60±0.01	59.18±0.09	20.17±0.23	15.03±0.04	28.62±0.62
PAAG	2782.70±9.63	17.84±0.01	18.05±0.02	2.34±0.02	45.83±0.03	0.09±0.01	28.39±0.07	0.07±0.03	25.38±0.01	0.10±0.03
Pinal	308.97±0.68	5.81±0.02	5.78±0.02	12.83±0.13	58.26±0.16	4.73±0.06	75.25±0.19	68.93±0.33	10.96±0.10	58.41±0.38
ProDVa	415.64±7.40	7.63±0.09	8.83±0.17	1.92±0.05	35.65±0.15	2.81±0.13	76.84±0.17	76.27±0.59	8.67±0.05	67.65±0.43

F.3 COMPLETE RESULTS FOR THE KEYWORD-GUIDED TASK

Results on keyword-guided task are in Table 9, Table 10 and Table 11. We mark the top three models, with deeper colors indicating superior performance.

Table 7: Benchmark results of Language Alignment for the description-guided task.

Models	Model-based Alignment		Retrieval-based Alignment								
	ProTrek Score ↑	EvoLlama Score ↑	Soft(4) ↑	Soft(10) ↑	Soft(20) ↑	Normal(4) ↑	Normal(10) ↑	Normal(20) ↑	Hard(4) ↑	Hard(10) ↑	Hard(20) ↑
Natural	27.00	60.33	98.50	97.17	96.09	93.72	89.01	85.11	43.23	33.20	29.78
Random(U)	1.03±0.04	36.22±0.07	28.09±0.83	12.62±0.29	6.94±0.20	28.97±0.38	12.83±0.52	7.16±0.29	25.95±0.52	10.57±0.40	5.38±0.39
Random(E)	1.04±0.05	34.11±0.10	28.35±0.62	12.83±0.51	6.73±0.66	28.97±0.62	12.59±0.29	6.84±0.39	25.79±0.45	10.39±0.67	5.46±0.45
ProteinDT	1.20±0.06	40.57±0.05	42.91±0.68	24.97±1.33	16.77±1.16	34.58±0.99	16.56±0.44	9.43±0.33	25.09±1.17	10.38±0.55	5.01±0.43
Chroma	2.10±0.02	40.10±0.23	29.54±0.59	13.43±0.18	7.41±0.22	29.63±0.58	13.26±0.50	7.44±0.21	25.51±0.47	10.68±0.43	5.73±0.19
PAAG	1.29±0.04	34.39±0.18	33.33±0.25	15.24±0.33	8.27±0.20	29.63±0.70	12.83±0.14	6.87±0.17	25.19±0.65	10.13±0.34	4.96±0.23
Pinal	17.50±0.09	53.40±0.31	82.42±0.45	74.44±0.63	69.99±0.67	71.69±0.59	63.53±0.24	58.43±0.48	29.51±0.31	17.89±0.26	14.17±0.27
ProDVa	17.40±0.06	51.19±0.17	85.64±0.06	77.37±0.44	72.75±0.61	77.52±0.28	66.67±0.37	59.03±0.44	27.84±0.71	15.77±0.29	12.11±0.28

Table 8: Benchmark results of Similarity, Novelty, and Diversity for the description-guided task.

Models	Similarity					Novelty				Diversity	
	GT-Identity ↑	GT-TMScore ↑	ESM-F1 ↑	ESM-Precision ↑	ESM-Recall ↑	Seq_{Easy} ↑	Seq_{Hard} ↑	$Struct_{Easy}$ ↑	$Struct_{Hard}$ ↑	Sseq ↑	Struct ↑
Natural	100.00	100.00	100.00	100.00	100.00	36.11	4.90	38.51	13.56	-	-
Random(U)	0.37±0.03	16.95±0.03	71.06±0.02	81.66±0.02	63.46±0.02	98.77±0.03	58.14±0.07	96.82±0.03	77.64±0.12	97.01	81.59
Random(E)	0.23±0.04	17.10±0.00	71.95±0.02	82.51±0.02	64.35±0.02	98.45±0.01	60.19±0.14	96.25±0.04	76.82±0.10	99.56	81.45
ProteinDT	0.18±0.02	13.94±0.03	72.80±0.05	81.44±0.03	66.38±0.05	96.92±0.12	70.74±0.07	94.68±0.02	71.16±0.08	99.23	83.67
Chroma	0.22±0.04	17.93±0.02	72.82±0.02	80.22±0.03	67.06±0.01	97.28±0.02	58.68±0.09	80.99±0.04	51.06±0.21	96.13	79.9
PAAG	0.17±0.02	14.63±0.03	73.26±0.03	83.10±0.02	66.04±0.03	98.90±0.02	63.64±0.09	96.44±0.03	77.34±0.12	99.15	82.16
Pinal	18.65±0.15	23.75±0.14	76.63±0.06	77.74±0.08	75.99±0.06	55.55±0.19	43.82±0.22	40.07±0.33	17.23±0.23	82.96	72.73
ProDVa	21.48±0.15	20.03±0.11	75.23±0.01	77.01±0.05	74.11±0.02	38.23±0.31	14.64±0.23	56.18±23.36	36.31±33.02	83.29	36.92

Table 9: Benchmark results of Plausibility and Foldability for the keyword-guided task.

Models	Perplexity			Repetitiveness			Foldability			
	PPL-ProtGPT2 ↓	PPL-ProGen ↓	PPL-RITA ↓	Repeat ↓	Rep-2 ↓	Rep-5 ↓	pLDDT ↑	% > 70 ↑	PAE ↓	% < 10 ↑
guided with GO keywords										
Natural	554.35	9.17	8.89	2.17	44.43	0.43	76.92	72.44	10.54	54.69
Random(U)	2473.84±10.48	21.74±0.01	22.18±0.02	0.72±0.03	35.52±0.04	0.01±0.00	23.20±0.02	0.10±0.08	24.56±0.01	0.19±0.08
Random(E)	3096.13±18.42	18.68±0.01	19.05±0.01	1.14±0.04	41.45±0.07	0.02±0.00	25.99±0.12	0.05±0.08	24.47±0.03	0.24±0.08
ProteoGAN	2708.39±32.50	18.03±0.01	18.31±0.02	2.50±0.05	42.73±0.86	0.03±0.00	28.72±0.43	0.06±0.10	24.67±0.17	0.12±0.20
CFP-Gen	187.72±9.71	5.16±0.03	4.65±0.02	12.67±0.79	59.67±0.83	13.82±0.74	73.38±0.26	65.65±1.11	14.61±0.27	35.20±1.76
ProteinDT	1531.76±17.19	12.23±0.06	12.29±0.06	7.98±0.51	64.01±0.25	3.32±0.38	40.35±0.30	1.15±0.00	25.57±0.03	0.00±0.00
Chroma	1354.61±4.81	12.18±0.03	12.40±0.03	2.71±0.05	55.09±0.12	0.67±0.03	59.27±0.20	22.17±0.65	15.00±0.10	30.93±0.08
PAAG	2650.36±11.01	18.08±0.02	18.38±0.02	2.48±0.20	39.23±0.05	0.05±0.01	31.47±0.10	0.34±0.36	23.88±0.05	0.24±0.08
Pinal	414.26±77.15	6.85±0.59	6.89±0.64	14.13±2.58	59.84±4.37	4.85±1.63	72.58±5.55	62.10±14.24	11.79±2.52	52.19±12.03
ProDVa	486.77±9.51	11.16±0.29	18.71±0.77	1.87±0.07	22.04±0.09	0.88±0.05	74.73±0.24	68.40±0.38	6.11±0.02	84.90±0.46
guided with IPR keywords										
Natural	611.99	9.73	9.47	2.23	44.05	0.48	75.77	68.85	11.13	50.92
Random(U)	2475.07±10.83	21.76±0.02	22.21±0.02	0.69±0.06	35.18±0.07	0.01±0.00	23.40±0.07	0.08±0.07	24.42±0.02	0.11±0.00
Random(E)	3104.89±24.07	18.67±0.02	19.05±0.02	1.16±0.02	40.91±0.08	0.02±0.01	26.29±0.18	0.08±0.07	24.34±0.03	0.19±0.07
ESM3	330.44±9.90	6.33±0.07	6.59±0.07	28.13±0.24	68.98±0.42	21.11±0.47	60.90±0.77	32.93±2.43	16.73±0.27	22.68±1.83
CFP-Gen	135.57±4.51	4.94±0.12	5.03±0.11	11.86±0.29	59.17±0.57	13.57±0.88	76.36±0.35	72.52±1.45	12.54±0.26	47.23±2.51
ProteinDT	1506.64±5.70	11.87±0.02	11.93±0.02	10.02±0.38	65.68±0.30	5.83±0.39	37.59±0.15	0.04±0.07	26.19±0.03	0.00±0.00
Chroma	1336.19±7.55	12.17±0.01	12.39±0.02	2.60±0.08	54.53±0.08	0.54±0.03	59.76±0.26	23.75±1.35	14.67±0.05	31.38±0.75
PAAG	2748.12±25.25	17.85±0.04	18.06±0.03	2.32±0.11	44.78±0.06	0.08±0.01	30.89±0.03	0.11±0.11	24.98±0.02	0.19±0.13
Pinal	525.38±80.49	8.12±0.45	8.22±0.47	16.73±1.96	59.97±3.55	6.32±1.28	65.69±5.42	44.90±12.43	14.10±2.19	36.13±9.57
ProDVa	574.60±5.52	12.47±0.77	19.07±1.06	1.99±0.02	21.64±0.11	1.51±0.12	72.80±0.48	60.65±0.65	6.86±0.10	79.92±1.03
guided with IPR&GO keywords										
Natural	534.49	8.96	8.66	2.16	45.01	0.44	77.17	73.15	10.48	54.9
Random(U)	2482.06±21.06	21.72±0.01	22.14±0.01	0.73±0.05	36.12±0.06	0.01±0.00	22.85±0.10	0.00±0.00	24.72±0.01	0.00±0.00
Random(E)	3120.95±10.54	18.68±0.02	19.03±0.01	1.14±0.03	42.08±0.02	0.02±0.00	25.60±0.06	0.00±0.00	24.59±0.02	0.00±0.00
CFP-Gen	163.51±6.62	5.23±0.04	5.23±0.07	13.14±1.18	59.86±0.53	14.17±0.32	72.70±1.07	60.90±1.11	14.45±0.28	42.69±1.76
ProteinDT	1697.89±9.15	12.81±0.05	12.87±0.05	6.81±0.17	63.58±0.36	2.91±0.21	36.46±0.31	0.20±0.17	25.75±0.06	0.00±0.00
Chroma	1360.90±5.58	12.19±0.04	12.40±0.04	2.53±0.12	55.65±0.16	0.56±0.05	58.71±0.59	19.29±2.67	15.33±0.30	29.72±1.34
PAAG	2807.41±13.89	17.80±0.01	17.98±0.02	2.32±0.04	47.64±0.24	0.09±0.01	30.05±0.27	0.00±0.00	25.69±0.10	0.00±0.00
Pinal	442.23±69.42	7.39±0.49	7.49±0.51	16.22±2.42	59.71±3.71	6.07±1.49	69.32±5.10	53.56±12.44	12.97±2.18	42.53±10.69
ProDVa	500.40±7.93	10.48±0.07	13.61±0.95	2.61±0.43	28.73±0.63	3.89±0.68	74.26±0.27	67.46±1.49	8.06±0.07	72.16±0.23
GPT-5	769.53±13.47	10.87±0.14	10.99±0.13	4.63±0.14	64.60±0.26	15.41±0.61	32.12±0.11	1.93±0.45	24.55±0.02	0.69±0.45

G Extended Experiments

G.1 Can large language models handle protein design from function?

In addition to the aforementioned baselines, we aim to investigate whether large language models such as ChatGPT can effectively perform the task. Using the API provided by OpenAI, we accessed the GPT-5 model and employed it with a specialized system prompt and the function description as user input, as illustrated in Table 12.

As shown in Table 11&10&11, GPT-5 outperforms PAAG, Chroma, and ProteinDT across most language alignment metrics, though it still falls markedly short of CFP-Gen, Pinal, and ProDVa. It also achieves respectable results in both perplexity and repetitiveness, suggesting that large language

Table 10: Benchmark results of Language Alignment for the keyword-guided task.

Models	Model-based Alignment			Retrieval-based Alignment								
	ProTrek Score ↑	IPR Recovery ↑	GO Recovery ↑	Soft(4) ↑	Soft(10) ↑	Soft(20) ↑	Normal(4) ↑	Normal(10) ↑	Normal(20) ↑	Hard(4) ↑	Hard(10) ↑	Hard(20) ↑
guided with GO keywords												
Natural	21.6	100.0	100.0	94.52	92.78	89.75	87.59	77.49	69.41	37.23	28.72	26.7
Random(U)	4.29±0.04	0.00±0.00	20.79±0.37	30.06±1.33	14.00±4.45	7.50±2.57	26.89±1.83	10.00±0.08	5.05±1.44	30.01±3.21	9.72±0.46	5.05±0.29
Random(E)	3.44±0.01	0.00±0.00	11.71±1.48	29.87±0.50	12.41±2.18	6.69±1.04	27.08±0.30	11.06±0.55	5.96±0.98	28.62±2.22	11.98±0.66	5.87±0.79
ProteoGAN	4.42±0.02	0.00±0.00	14.99±1.36	38.64±2.86	19.84±2.14	10.27±1.11	32.70±1.42	13.84±0.69	8.13±0.52	27.16±1.67	10.90±1.05	5.65±0.53
CFP-Gen	10.03±0.33	9.67±1.07	18.98±1.14	66.24±2.17	57.40±1.85	47.39±3.10	53.46±2.49	38.87±2.00	30.09±0.92	27.80±3.72	13.37±0.74	8.15±0.42
ProteinDT	1.70±0.21	0.03±0.05	18.52±0.71	37.52±6.22	16.59±1.25	8.90±0.36	34.05±1.91	15.39±1.42	8.51±0.88	27.08±1.79	12.51±1.23	6.30±0.73
Chroma	1.84±0.03	0.23±0.05	16.33±2.36	32.13±0.79	13.80±2.03	7.74±0.87	27.90±0.96	12.07±1.06	5.92±0.38	27.99±0.43	12.22±0.92	5.15±0.82
PAAG	4.38±0.17	0.00±0.00	21.66±2.71	38.58±1.75	20.49±1.59	11.16±1.42	34.10±0.92	16.45±0.76	8.85±0.79	31.36±2.33	12.07±0.46	6.30±1.21
Pinal	12.69±1.42	19.26±1.90	22.76±1.78	73.98±4.37	61.90±8.29	56.66±7.96	61.52±4.86	49.93±3.82	42.95±3.35	35.88±4.96	21.26±2.48	17.89±2.27
ProDVa	14.42±0.07	20.22±0.14	30.24±0.75	86.48±1.78	71.38±5.34	55.80±0.60	66.43±0.58	52.38±0.66	45.07±0.30	35.93±3.38	21.45±2.75	18.13±0.96
guided with IPR keywords												
Natural	25.29	100.0	100.0	98.51	96.67	95.17	91.72	83.22	75.75	40.8	32.76	30.69
Random(U)	7.53±0.07	0.00±0.00	25.75±3.21	27.70±3.13	10.84±1.76	6.44±0.75	26.44±1.78	11.72±0.11	6.21±0.64	28.05±1.44	9.85±0.63	4.87±0.24
Random(E)	6.11±0.08	0.00±0.00	13.06±1.17	27.62±2.25	13.18±1.93	8.01±1.29	28.51±2.22	12.80±0.13	6.74±0.93	26.36±0.65	11.34±1.53	5.48±0.27
ESM3	6.22±0.18	20.17±0.86	15.43±2.74	55.41±7.66	37.77±1.88	31.64±1.72	48.31±1.52	33.01±1.24	26.47±1.63	29.82±0.88	14.69±0.18	10.69±0.43
CFP-Gen	10.21±0.16	32.79±0.93	23.41±1.82	64.47±1.84	50.78±1.88	43.36±0.65	55.90±1.73	40.96±1.19	34.38±1.48	29.36±1.73	15.88±1.31	12.43±0.48
ProteinDT	3.85±0.03	0.08±0.05	20.76±1.46	40.38±3.85	22.68±1.32	13.83±0.98	34.44±0.29	16.13±1.04	9.23±0.58	26.82±2.94	10.92±1.66	5.63±1.41
Chroma	3.82±0.03	0.17±0.01	17.15±1.93	37.13±1.05	17.05±3.30	9.27±1.04	29.35±1.09	13.68±0.30	6.78±0.20	27.20±2.24	11.15±0.87	5.25±0.87
PAAG	5.98±0.17	0.08±0.05	13.85±1.02	32.07±3.27	14.41±2.50	9.58±1.78	30.69±2.26	14.37±1.85	7.78±0.86	26.90±0.92	11.46±2.17	6.05±0.87
Pinal	14.38±1.38	25.63±3.95	15.93±1.49	80.96±7.50	70.88±9.31	64.21±10.47	71.00±4.87	57.59±5.60	48.24±6.90	31.88±1.90	20.31±1.36	16.59±1.90
ProDVa	15.19±0.19	24.58±0.69	26.59±0.40	80.15±2.25	64.90±3.11	57.05±4.15	65.29±0.64	51.99±0.93	44.44±0.88	33.26±1.24	20.73±1.50	15.90±0.54
guided with IPR&GO keywords												
Natural	27.36	100.0	100.0	99.55	99.11	98.96	93.62	87.39	79.82	45.85	36.65	34.27
Random(U)	4.84±0.09	0.00±0.00	25.38±3.32	29.77±1.88	12.86±2.64	5.64±1.27	26.36±1.34	9.45±1.44	4.50±1.84	26.41±1.12	9.69±0.45	4.75±1.07
Random(E)	3.72±0.09	0.00±0.00	14.67±4.14	30.42±2.20	12.12±0.62	5.93±0.90	27.00±1.46	11.28±1.04	6.28±1.30	26.56±1.46	10.29±1.30	5.24±0.31
CFP-Gen	11.68±0.15	35.21±0.30	23.31±2.49	73.97±1.60	63.46±2.00	59.23±1.54	57.95±1.94	45.77±1.92	35.51±3.11	29.36±1.46	15.00±0.00	11.54±0.77
ProteinDT	3.06±0.11	0.36±0.03	15.92±1.14	47.08±1.20	29.13±1.43	19.63±1.63	38.72±1.43	19.29±0.68	9.99±0.97	27.45±2.32	10.48±0.69	5.39±1.09
Chroma	2.19±0.12	0.16±0.06	14.12±1.86	32.29±2.15	14.34±2.12	8.56±0.48	28.83±0.69	11.67±1.65	6.38±0.53	27.00±2.50	11.67±0.73	5.79±0.74
PAAG	4.66±0.12	0.02±0.02	9.77±0.90	28.24±3.82	14.14±3.71	7.22±2.06	28.19±1.67	11.82±0.82	6.03±0.60	29.08±1.80	12.17±1.55	6.08±0.51
Pinal	15.26±1.27	33.08±3.75	21.64±0.47	82.34±3.25	73.10±4.55	68.55±4.53	72.50±3.87	60.88±2.79	52.62±3.12	34.47±1.34	21.76±1.12	18.50±1.24
ProDVa	16.78±0.12	30.95±0.56	25.24±0.45	82.54±1.19	74.18±1.36	69.24±1.63	71.51±0.93	61.23±1.50	52.97±2.43	33.88±1.79	22.45±1.09	19.93±1.09
GPT-5	9.13±0.10	6.64±0.31	29.43±1.02	77.50±2.08	63.60±2.31	54.35±1.57	61.08±2.41	42.78±2.27	29.97±1.51	26.61±0.45	12.81±0.67	7.52±0.17

Table 11: Benchmark results of Similarity, Novelty and Diversity for the keyword-guided task.

Models	Similarity					Novelty				Diversity	
	GT-Identity ↑	GT-TMScore ↑	ESM-F1 ↑	ESM-Precision ↑	ESM-Recall ↑	Seq_{Easy} ↑	Seq_{Hard} ↑	$Struct_{Easy}$ ↑	$Struct_{Hard}$ ↑	Seq ↑	Struct ↑
guided with GO keywords											
Natural	100.0	100.0	100.0	100.0	100.0	44.34	4.07	56.96	18.15	-	-
Random(U)	0.84±0.23	16.76±0.05	73.37±0.04	82.81±0.05	66.40±0.03	98.66±0.07	58.04±0.69	96.54±0.11	76.75±0.20	94.29	81.56
Random(E)	0.63±0.16	17.03±0.03	74.24±0.04	83.70±0.05	67.24±0.03	98.44±0.04	60.28±0.09	95.85±0.15	75.92±0.19	98.65	81.54
ProteoGAN	0.28±0.07	14.75±0.24	74.25±0.07	84.37±0.10	66.84±0.05	99.13±0.06	65.24±0.27	96.19±0.17	75.82±0.29	98.94	84.37
CFP-Gen	2.30±0.25	13.98±0.26	67.52±0.17	68.36±0.26	67.46±0.30	59.60±0.61	47.85±0.95	54.07±1.87	28.28±1.47	85.14	81.76
ProteinDT	0.20±0.12	12.67±0.05	74.83±0.02	82.67±0.10	68.93±0.05	99.28±0.06	75.41±0.30	96.29±0.05	74.62±0.63	99.7	84.53
Chroma	0.38±0.07	17.67±0.06	74.29±0.02	80.80±0.02	69.15±0.02	97.44±0.16	59.35±0.58	80.22±0.45	50.88±0.84	93.7	79.79
PAAG	0.16±0.05	16.22±0.18	75.40±0.05	84.22±0.07	68.77±0.05	98.80±0.03	62.36±0.47	95.20±0.27	73.36±0.18	98.57	81.73
Pinal	5.35±0.35	15.84±0.47	71.36±0.66	72.24±1.85	71.06±0.46	61.98±8.67	46.06±5.23	46.42±9.10	19.27±6.43	87.61	79.0
ProDVa	9.07±0.14	20.25±0.13	72.54±0.10	75.81±0.06	70.25±0.12	48.37±0.72	25.02±0.81	62.12±0.19	32.72±0.02	98.17	35.76
guided with IPR keywords											
Natural	100.0	100.0	100.0	100.0	100.0	44.92	4.47	59.46	20.09	-	-
Random(U)	0.88±0.03	16.69±0.04	73.97±0.04	82.79±0.02	67.38±0.05	98.70±0.06	57.21±0.77	96.37±0.05	76.47±0.35	94.44	81.56
Random(E)	0.70±0.15	16.85±0.12	74.79±0.03	83.65±0.03	68.16±0.04	98.44±0.08	59.57±0.81	95.62±0.14	75.05±0.48	98.58	81.46
ESM3	4.43±0.22	21.30±0.28	72.22±0.09	75.06±0.13	69.96±0.25	85.30±0.68	71.87±1.23	73.80±1.08	37.56±0.41	91.41	76.79
CFP-Gen	7.75±0.16	16.73±0.42	66.82±0.17	68.61±0.13	65.74±0.25	63.79±0.11	49.46±0.57	50.44±1.32	23.15±0.97	85.31	82.08
ProteinDT	0.13±0.02	12.38±0.06	75.23±0.04	82.13±0.08	70.07±0.02	99.08±0.13	73.57±0.65	96.64±0.12	76.03±0.19	99.71	84.81
Chroma	0.38±0.13	17.45±0.06	74.81±0.02	80.85±0.03	70.00±0.01	97.35±0.13	59.25±0.27	80.19±0.68	50.77±1.26	94.06	79.88
PAAG	0.26±0.07	14.37±0.07	76.19±0.01	84.43±0.03	69.93±0.03	98.93±0.07	64.71±0.70	96.66±0.16	79.23±0.22	99.16	81.48
Pinal	6.70±0.90	17.23±0.57	74.14±0.28	76.19±1.29	72.59±0.62	74.01±6.01	51.61±3.45	60.24±6.98	27.00±6.19	87.02	80.38
ProDVa	7.39±0.08	20.75±0.36	73.31±0.09	76.62±0.17	70.99±0.09	51.80±1.34	28.86±1.37	65.58±0.77	31.26±1.57	95.02	45.6
guided with IPR&GO keywords											
Natural	100.0	100.0	100.0	100.0	100.0	43.23	3.89	56.17	17.72	-	-
Random(U)	0.85±0.34	16.73±0.10	73.09±0.02	82.66±0.03	66.02±0.01	98.85±0.03	57.48±0.13	96.58±0.23	77.18±0.45	94.24	81.73
Random(E)	0.69±0.04	16.89±0.05	73.97±0.06	83.59±0.05	66.85±0.05	98.54±0.03	60.06±0.40	95.79±0.02	75.98±0.28	98.39	81.57
CFP-Gen	8.00±0.30	16.12±0.13	66.07±0.27	68.60±0.12	64.41±0.40	65.78±1.27	54.72±1.90	52.97±1.96	28.89±0.77	80.61	81.91
ProteinDT	0.29±0.14	13.08±0.03	74.91±0.02	82.88±0.06	68.91±0.06	98.75±0.06	71.44±0.50	96.53±0.05	75.73±0.21	99.39	84.18
Chroma	0.29±0.08	17.51±0.10	74.09±0.06	80.76±0.07	68.84±0.05	97.49±0.07	59.36±0.05	80.76±0.50	51.45±1.07	94.27	79.97
PAAG	0.20±0.08	13.77±0.07	75.52±0.05	84.45±0.05	68.81±0.05	99.12±0.06	65.07±0.68	97.22±0.09	81.53±0.18	99.22	81.51
Pinal	9.41±1.27	18.11±0.94	74.00±0.38	75.46±1.50	73.01±0.68	66.95±6.76	49.03±4.31	54.41±7.66	22.43±6.35	85.2	78.2
ProDVa	10.08±0.36	19.36±0.26	73.11±0.15	75.47±0.22	71.48±0.10	46.43±0.51	21.97±0.59	60.12±0.63	24.20±0.41	91.94	52.18
GPT5	2.99±0.03	14.63±0.07	75.14±0.06	81.36±0.07	70.30±0.07	91.42±0.06	54.87±0.36	92.44±0.18	66.58±0.31	87.66	80.65

models such as GPT-5 possess a solid grasp of protein sequences as well as textual data. However, GPT-5's performance in Foldability ranks near the bottom among all models, revealing significant limitations in its understanding of protein structures. In terms of similarity, novelty, and diversity, GPT-5 performs on par with ProteinDT, Chroma, and PAAG, metrics that often do not directly indicate the quality of the designed sequences.

G.2 How does the number of keywords impact the design capability?

Table 12: Settings of GPT-5

temperature	1.0
top_p	0.1
system prompt	You are an expert in designing novel functional proteins tailored to user requirements. Do not explain your answer; directly provide an amino acid sequence.
user prompt	$\{The\ functional\ description\ of\ the\ protein\}$

Designing proteins that satisfy a greater number of functional keywords is inherently more challenging. Accordingly, we compare performance across tasks of varying difficulty. We first calculated the number of InterPro entries and Gene Ontology (Molecular Function) terms associated with each protein in SwissProt. Using the median frequency as a threshold, we classified each keyword-guided task into **Easy** and **Hard** sub-tasks. The median frequency was 2 for Gene Ontology (Molecular Function) terms, 7 for InterPro entries, and 9 for combined IPR&GO keywords. Overall, GPT-5 outperforms ProteinDT, PAAG, and Chroma in protein sequence comprehension and language alignment, yet it remains markedly limited in its understanding of protein structures.

Table 13: Results for Plausibility, Foldability, and Alignment, categorized by keyword count.

Model	PPL-ProGen			pLDDT			ProTrek Score		
	Easy	Hard	Overall	Easy	Hard	Overall	Easy	Hard	Overall
guided with GO keywords (The samples with >2 GO terms are treated as Hard.)									
Natural	9.99	7.05	9.17	75.64	80.21	76.92	21.05	23.00	21.60
ProteoGAN	18.00 ± 0.02	18.08 ± 0.03	18.03 ± 0.01	28.89 ± 0.41	28.35 ± 0.47	28.72 ± 0.43	4.98 ± 0.14	3.28 ± 0.24	4.42 ± 0.02
CFP-Gen	5.05 ± 0.01	5.48 ± 0.13	5.16 ± 0.03	75.95 ± 0.27	65.90 ± 0.24	73.38 ± 0.26	10.82 ± 0.44	7.75 ± 0.08	10.03 ± 0.33
ProDVa	11.59 ± 0.28	10.08 ± 0.47	11.16 ± 0.29	73.70 ± 0.30	77.40 ± 0.09	74.73 ± 0.24	14.71 ± 0.08	13.68 ± 0.07	14.42 ± 0.07
guided with IPR keywords (The samples with >7 IPR entries are treated as Hard.)									
Natural	11.23	6.87	9.73	72.46	82.04	75.77	25.82	24.29	25.29
ESM3	26.82 ± 0.10	5.30 ± 0.01	6.33 ± 0.07	59.87 ± 0.54	63.08 ± 1.27	60.90 ± 0.77	6.21 ± 0.12	6.24 ± 0.69	6.22 ± 0.18
CFP-Gen	5.04 ± 0.13	4.65 ± 0.13	4.94 ± 0.12	76.98 ± 0.14	74.57 ± 0.93	76.36 ± 0.35	8.86 ± 0.23	14.03 ± 0.42	10.21 ± 0.16
ProDVa	14.21 ± 1.05	9.18 ± 0.30	12.47 ± 0.77	70.38 ± 0.46	77.41 ± 0.71	72.80 ± 0.48	14.25 ± 0.25	16.98 ± 0.10	15.19 ± 0.19
guided with IPR&GO keywords (The samples with >9 keywords are treated as Hard.)									
Natural	10.64	6.71	8.96	73.61	81.96	77.17	28.59	25.70	27.36
CFP-Gen	5.34 ± 0.19	4.96 ± 0.31	5.23 ± 0.04	71.60 ± 1.31	75.26 ± 0.55	72.70 ± 1.07	9.52 ± 0.09	16.73 ± 0.34	11.68 ± 0.15
ProDVa	12.44 ± 0.23	7.87 ± 0.17	10.48 ± 0.07	73.56 ± 0.23	75.19 ± 0.71	74.26 ± 0.27	15.40 ± 0.08	18.62 ± 0.24	16.78 ± 0.12

Table 14: Results for Similarity, Novelty and Diversity, categorized by keyword count.

Model	GT-Identity			Novelty-Seq$_{\text{Hard}}$			Diversity-Seq		
	Easy	Hard	Overall	Easy	Hard	Overall	Easy	Hard	Overall
guided with GO keywords (The samples with >2 GO terms are treated as Hard.)									
Natural	100.00	100.00	100.00	4.25	3.59	4.07	-	-	-
ProteoGAN	0.12 ± 0.04	0.60 ± 0.25	0.28 ± 0.07	65.36 ± 0.57	64.97 ± 1.49	65.24 ± 0.27	99.42	97.94	98.94
CFP-Gen	1.78 ± 0.14	3.80 ± 0.66	2.30 ± 0.25	41.65 ± 0.78	65.89 ± 1.68	47.85 ± 0.95	84.42	87.22	85.14
ProDVa	7.73 ± 0.27	12.50 ± 0.81	9.07 ± 0.14	27.19 ± 1.22	19.45 ± 0.63	25.02 ± 0.81	98.57	97.13	98.17
guided with IPR keywords (The samples with >7 IPR entries are treated as Hard.)									
Natural	100.00	100.00	100.00	4.75	3.94	4.47	-	-	-
ESM3	3.06 ± 0.09	7.30 ± 0.84	4.43 ± 0.22	73.24 ± 1.30	68.99 ± 2.57	71.87 ± 1.23	93.14	87.78	91.41
CFP-Gen	4.19 ± 0.18	17.87 ± 1.11	7.75 ± 0.16	50.42 ± 0.68	46.73 ± 0.27	49.46 ± 0.57	87.98	77.72	85.31
ProDVa	3.64 ± 0.27	14.49 ± 0.69	7.39 ± 0.08	35.52 ± 2.01	16.27 ± 0.20	28.86 ± 1.37	97.20	90.89	95.02
guided with IPR&GO keywords (The samples with >9 keywords are treated as Hard.)									
Natural	100.00	100.00	100.00	4.10	3.61	3.89	-	-	-
CFP-Gen	3.55 ± 0.33	18.37 ± 0.86	8.00 ± 0.30	57.52 ± 2.31	48.21 ± 0.94	54.72 ± 1.90	83.20	74.57	80.61
ProDVa	5.40 ± 0.39	16.35 ± 0.40	10.08 ± 0.36	28.79 ± 1.56	12.83 ± 0.72	21.97 ± 0.59	96.35	86.02	91.94

As shown in Table 13&14, We selected representative baselines for three keyword-guided subtasks, choosing one metric from each of six dimensions. The evaluation results were categorized as Easy, Hard, and Overall (without distinction). Based on the results for Natural Protein, plausibility, foldability, and similarity improve as the number of keywords increases, whereas language alignment, novelty, and diversity decline. Most baselines follow trends similar to those observed for Natural Protein, though some exceptions occur. For example, CFP-Gen consistently attains optimal and stable plausibility across both Hard and Easy samples. In contrast, ProTrek Score

exhibits divergent trends across subtasks, which we attribute to dataset imbalance resulting from the limited volume of test data.

G.3 What is the optimal evaluation strategy when using PDFBench?

As noted in Section 4.1, the Perplexity, used to assess sequence plausibility, can to some extent capture two metrics of Foldability, the latter typically requiring extremely more computational time. In this section, we evaluate the computational costs of proposed metrics and propose a practical evaluation scheme.

Here, we randomly selected 50 and 500 samples from *SwissTest*and performed evaluations on a single machine equipped with 4 NVIDIA GeForce RTX3090 and one 48-core CPU. For Perplexity, we only evaluated the PPL-ProGen2; For Repetitiveness, we computed only the Repeat; For Foldability, both pLDDT and PAE metrics can be obtained simultaneously in a single inference of ESMFold; For Novelty and Diversity, we considered computing both Sequence and Structure metrics concurrently.

Table 15: Computation time for all metrics. The values below represent the mean time required to evaluate 50 samples.

Num of Samples	Plausibility		Foldability	Language Alignment				Similarity			Novelty	Diversity
	Perplexity	Repetitiveness		ProTrek Score	IPR Recovery	Retrieval Accuracy	GO Recovery	ESMScore	GT-Identity	GT-TMScore		
50	39s	2s	9m34s	32s	3s	38s	65s	11s	18s	18s	19s	21s
500	23s	3s	61m50s	47s	29m39s	104s	103s	24s	91s	35s	38s	18s

As shown in Table 15, the computational cost of Foldability is 160 times greater than that of Perplexity, as previously noted. The IPR Recovery is time-intensive, primarily because it relies on *InterProScan*—a Java-based retrieval tool that cannot be accelerated using GPUs. In summary, we recommend using metrics other than Foldability and IPR Recovery for the initial screening, followed by these two metrics in a subsequent evaluation round. Furthermore, it should be noted that all metrics, except Repetitiveness, IPR Recovery, and GT-TMScore, depend on GPU processing. Therefore, the GPU requirements of PDFBenchshould not be underestimated.

H Limitations

In this study, we conducted a fair and comprehensive benchmark of two tasks, 8 models, including ProteinDT, Chroma, PAAG, Pinal, ProDVa, ProteoGAN, ESM3, CFP-Gen across all 16 metrics. For keyword-guided tasks, to ensure a fair comparison, we restricted the evaluation dataset to 1,057 proteins released between January 1, 2025, and August 25, 2025. It should be noted that this limited dataset size may affect the reliability of the evaluation results.