PDB-Struct: A Comprehensive Benchmark for Structure-based Protein Design

Chuanrui Wang[1,2] **Bozitao Zhong** [1,2] **Zuobai Zhang**[1,2]
Narendra Chaudhary[3] **Sanchit Misra**[3] **Jian Tang**[1,4,5 †]
[1] Mila - Québec AI Institute [2] Université de Montréal [3] Intel Parallel Computing Lab
[4] HEC Montréal [5] CIFAR AI Chair
{chuanrui.wang, bozitao.zhong, zuobai.zhang}@mila.quebec,
{narendra.chaudhary, sanchit.misra}@intel.com, jian.tang@hec.ca

Abstract

Structure-based protein design has attracted increasing interest, with numerous methods being introduced in recent years. However, a universally accepted method for evaluation has not been established, since the wet-lab validation can be overly time-consuming for the development of new algorithms, and the *in silico* validation with recovery and perplexity metrics is efficient but may not precisely reflect true foldability. To address this gap, we introduce two novel metrics: refoldability-based metric, which leverages high-accuracy protein structure prediction models as a proxy for wet lab experiments, and stability-based metric, which assesses whether models can assign high likelihoods to experimentally stable proteins. We curate datasets from high-quality CATH protein data, high-throughput *de novo* designed proteins, and mega-scale experimental mutagenesis experiments, and in doing so, present the **PDB-Struct** benchmark that evaluates both recent and previously uncompared protein design methods. Experimental results indicate that ByProt, ProteinMPNN, and ESM-IF perform exceptionally well on our benchmark, while ESM-Design and AF-Design fall short on the refoldability metric. We also show that while some methods exhibit high sequence recovery, they do not perform as well on our new benchmark. Our proposed benchmark paves the way for a fair and comprehensive evaluation of protein design methods in the future. Code is available at https://github.com/WANG-CR/PDB-Struct.

1 Introduction

Designing new proteins with desired properties is a crucial task in bioengineering [Huang et al., 2016]. It aids in developing therapies, crafting novel antibodies, and exploring the uncharted realm of proteins beyond those found in nature. Structure-based protein design has emerged as the predominant approach for de novo protein design, owing to its versatile application across proteins with well-defined structures. In recent years, the integration of deep learning has enhanced the capabilities of structure-based protein design, yielding notable results [Ingraham et al., 2019, Jing et al., 2020, Dauparas et al., 2022, Zheng et al., 2023]. This progress is evidenced by the creation of miniproteins specifically engineered to bind particular targets [Cao et al., 2022], enzymes fine-tuned for new substrates [Yeh et al., 2023], and innovative antibodies [Luo et al., 2022, Shi et al., 2022].

Benchmarking these methods is of paramount importance; however, current benchmarks have limitations. While experimental validation of generated sequences is costly [Ladd et al., 1977, Bai et al., 2015, Dauparas et al., 2022], the *in silico* proxy, which calculates sequence recovery and perplexity [Jing et al., 2020] on a test set of natural proteins, serves as an effective replacement but

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

may not accurately reflect real-world foldability. The recovery is calculated as the similarity between the designed sequence and the ground truth sequence, but high sequence similarity can not sufficiently imply the ability to fold into similar structures because even single mutations could cause a protein to misfold, such as Alzheimer's and cystic fibrosis [Cohen and Kelly, 2003, Qu et al., 1997]. The perplexity evaluates the uncertainty of a model's predictions by measuring the likelihood that the protein design model assigns to the ground truth sequence. However, the ground truth sequence only provide a point probability mass function instead of the true distribution landscape. Furthurmore, protein design methods compute pseudo-likelihoods based on varying assumptions, making these scores incomparable. For instance, one-shot prediction models [Gao et al., 2023b] compute it under the conditional independence assumption, whereas autoregressive models [Ingraham et al., 2019] do not use this assumption.

To address these limitations and complement the known metrics, we propose two novel metrics to benchmark structure-based protein design methods. The first metric, termed "refoldability", assesses the quality of designed sequences. This quality is determined by two factors: whether the designed sequences can fold into a stable structure and whether this stable structure is similar to the input structure. We utilize the atom-level protein structure prediction models [Jumper et al., 2021, Mirdita et al., 2022, Lin et al., 2022, Wu et al., 2022] to predict structures for the designed sequences, and calculate the TM score between the predicted structure and the input structure, as well as the pLDDT score reflecting the folding stability. The second metric, termed "stability-based metric", measures whether the the protein design methods can accurately estimate protein sequence landscape. We have curated datasets comprising one structure template and multiple sequences, derived from high-throughput *de novo* protein design and mutagenesis experiments [Rocklin et al., 2017, Tsuboyama et al., 2023]. Within these sequences, better protein design methods should assign higher probability to the sequences with higher experimental stability score, returning a higher stability-based Spearman correlation. Using these metrics and datasets, we evaluate the latest models as well as some previously unexamined ones. We have named this benchmark **PDB-Struct**, signifying a **struct**ure-based **p**rotein **d**esign **b**enchmark. Our contributions can be outlined as follows:

- We introduced two evaluation metrics based on carefully curated datasets.
- We ran experiments with popular structure-based protein design models and have established a fresh benchmark termed **PDB-Struct**.
- The proposed benchmark is the first that compares encoder-decoder based protein design methods together with structure-prediction based methods.
- By analysing the benchmark results, we outline the pros and cons of each protein design model, providing guidance for protein scientists when choosing a model.

2 Preliminaries

2.1 Problem Definition

Protein can be represented as a pair of amino acid sequence and structure $(\mathcal{S}, \mathcal{X})$, where $\mathcal{S} = [s_1, s_2, \cdots, s_n]$ denotes its sequence of n residues with $s_i \in \{1, ..., 20\}$ indicating the type of the i-th residue, and $\mathcal{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2 ..., \boldsymbol{x}_n] \in \mathbb{R}^{n\times 4\times 3}$ denotes its structure with $\boldsymbol{x}_i$ representing the Cartesian coordinates of the i-th residue's backbone atoms, including N, C-α, C and O. The challenge posed by the structure-based protein design is to elucidate an effective model θ capable of learning the underlying mapping from the provided structure data to the corresponding sequence distribution, and then generate novel sequences $\hat{\mathcal{S}} \sim p_\theta(\mathcal{S}|\mathcal{X})$.

2.2 Existing Methods

Encoder-Decoder Model Traditional methods encode 3D structure data using hand-crafted features or direct atom positions, typically employing MLPs [O'Connell et al., 2018, Li et al., 2014] and CNNs [Qi and Zhang, 2020, Anand et al., 2022]. Alternatively, viewing protein structure as a k-NN graph of amino acids retains spatial information, making GNNs a favored encoder. StructTrans employ graph-based self-attention modules in their encoder-decoder framework and decode in an autoregressive manner[Ingraham et al., 2019]. Further advancements have been made by GVP [Jing et al., 2020], ProteinMPNN [Dauparas et al., 2022] and ESM-Inverse Folding [Hsu et al., 2022],

both showcasing significant improvements. Some of the latest models suggest decoding residues conditionally independently given their structure, which accelerates the generation process without compromising sequence recovery[Gao et al., 2023b]. Moreover, inspired by the achievements in protein language modeling, ByProt introduced a structure adapter to incorporate ESM2 model[Lin et al., 2022], then decode as iterative refinement and boasts high sequence recovery[Zheng et al., 2023]. Works on graph-based encoder-decoder paradigms are emerging[Tan et al., 2022, Gao et al., 2023a, Mao et al., 2023], setting new benchmark in sequence recovery metric.

Structure Prediction based Model Models of this kind utilize pretrained structure prediction models or pretrained language models [Yang et al., 2020, Jumper et al., 2021] to compute an energy function, and then utilise different sampling strategies to generate samples. Wang et al. [2022] proposed to sample with thousands of gradient steps. Alternatively, Verkuil et al. [2022] proposed to perform Markov chain Monte Carlo sampling steps combined with simulated annealing, all to minimize the loss functions defined by protein structure prediction models and the structure condition. Similarly, hallucination methods aim to maximize the KL divergence between the predicted structures and a background distribution [Anishchenko et al., 2021, Hie et al., 2022]. However, it should be noted that the sampling process in these models tends to be slower than that in encoder-decoder models.

Diffusion-based Model Diffusion models [Ho et al., 2020] offer an alternative to generate samples through denoising, and they potentially offer advantages when learning from limited data [Zaidi et al., 2022]. Yi et al. [2023] performs denoising in the graph attribute space and achieves high sequence recovery. There are other models applying diffusion models in the discrete sequence space, such as EvoDiff [Alamdari et al., 2023] and, ProteinGenerator [Lisanza et al., 2023]. Other works, like Chroma [Ingraham et al., 2022] and RFDiffusion [Watson et al., 2022], apply denoising in the structure space. Since they have not released the code, or do not apply to structure-based protein sequence design, we are not evaluating them at the moment.

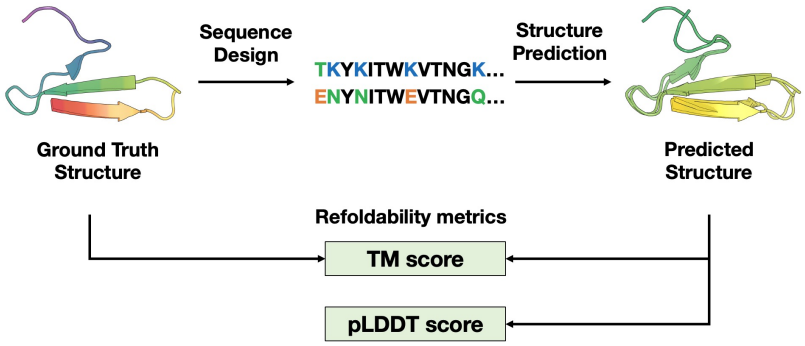


Figure 1: Pipeline of measuring refoldability metric.

3 Method

3.1 Evaluating the Designed Sequences with Refoldability-based Metrics

Motivation "Refoldability" is the natural metric that measures the quality of sequences designed based on structures. It evaluates the sequence quality on two aspects: whether the designed sequence can be expressed and fold stably, and whether they can refold into the input structure. Previous works have synthesized the proteins experimentally to assess refoldability [Dauparas et al., 2022, Verkuil et al., 2022]. However, the synthesis process [Mori and Barth, 1999], as well as the structure determination methods, such as X-ray crystallography and cryoEM [Ladd et al., 1977, Bai et al., 2015] are costly, hindering benchmarking across various design models. Previously, sequence recovery was proposed as an in silico benchmark [Ingraham et al., 2019]. While it's straightforward to calculate, there's no confirmed evidence that a high sequence similarity sufficiently implies a high

similarity between folded structures, or implies a good foldability in real world. For instance, even single mutations can cause a protein to misfold, leading to diseases such as Alzheimer's and cystic fibrosis [Cohen and Kelly, 2003, Qu et al., 1997]. Fortunately, due to advancements in high-accuracy protein structure prediction models, recent work [Wang et al., 2022] suggests leveraging them as an *in silico* proxy for actual structures. Adopting this idea, we propose to estimate the true refoldability with structure prediction models.

Metric The evaluation pipeline is shown in Figure. 1, where we generate multiple sequences with sequence design models given an input structure, and predict the structures for all the generated sequences. Firstly, to assess whether the generated sequences can respect the structure condition, we evaluate the agreement of the ground truth structure with the predicted structures using the TM-score [Zhang and Skolnick, 2005]. We refer this metric as **Ref-TM**. Furthurmore, to evaluate the folding stability of the generated sequences, we compute the mean value of the per-residue confidence estimate pLDDT predicted by the structure prediction models, refered as **Ref-pLDDT**. Previous research indicates that pLDDT serves as a reliable predictor of disorder [Tunyasuvunakool et al., 2021]. We employ AlphaFold2 [Jumper et al., 2021], OmegaFold [Wu et al., 2022], and ESMFold [Lin et al., 2023] as structure prediction models, which helps minimize deviations due to the choice of model.

It's important to highlight that, although Ref-TM metric and ScTM metric [Trippe et al., 2023] share a similar pipeline, they serve different purposes. The purpose of ScTM is to evaluate the quality of generated protein structures, treating both the protein design model and structure prediction model as oracles. In contrast, the foldability metric considers the inverse folding model as variable, while maintaining the input structure as a fixed ground truth derived from the test set.

Dataset We use the CATH4.2 40% non-redundant protein dataset [Orengo et al., 1997], and adopt the same data splitting based on CATH topology as StructTrans [Ingraham et al., 2019]. This results in 18024 protein single chains in the training set, 608 in the validation set, and 1120 in the test set. We furthur curated a small, high-quality test set from the original test set. After removing data points with unmeasured coordinates in the protein sequences, we randomly select one protein data from each CATH family and manually excluded proteins with extensive disordered regions, resulting in a final test set of 82 samples, with length ranges from 49 to 480 amino acids.

3.2 Evaluating the Estimated Likelihoods with Stability-based Metrics

Motivation Previous benchmark use perplexity as metric, which is the exponential of negative pseudo-log-likelihood. However, using and comparing perplexities introduces ambiguity due to several factors. First, the perplexity value is sensitive to changes in the sampling temperature. Using a protein design method with a high sampling temperature of 0.1, for example, could result in its perplexity exceeding that of a random sampling model based on residue frequency matrix, as demonstrated in Table. 1. Second, the computation of pseudo-log-likelihood differs among models, as shown in Table. 2. For example, PiFold assumes conditional independence of the residue types given the input structure, whereas ESM-IF does not make this assumption. Direct comparison between these methods, therefore, may not be entirely fair. Lastly, assigning high perplexity to the ground truth sequence does not imply that the protein design method construct the sequence distribution wrongly, since it is possible that the method has distributed a high probability mass function across many sequences which could fold into the given structure but are not present in the dataset.

Table 1: Perplexities on CATH test set.

Design method	Perplexity
Uniform	20.00
Natural frequencies*	18.32
ESM-IF($\tau = 1$)	4.24
ESM-IF($\tau = 0.1$)	3749.51

Table 2: Calculation of pseudo-log-likelihood.

Model Type[1]	Pseudo-log-likelihood $\mathcal{L}(\mathcal{S}\|\mathcal{X}, \theta)$
Autoregressive	$\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(s_i\|s_{<i}, \mathcal{X})$
One-Shot	$\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(s_i\|\mathcal{X})$
Refinement	$\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(s_i\|s_{-i}, \mathcal{X})$
MCMC	$-\lambda_p E_{projection}(\mathcal{X}\|\mathcal{S}) - \lambda_{LM} E_{LM}(\mathcal{S})$
Gradient Descent	$-\lambda_p E_{projection}(\mathcal{X}\|\mathcal{S}) - \lambda_{LM} E_{LM}(\mathcal{S})$

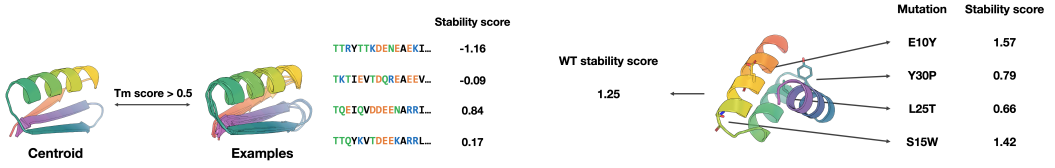
[1]The autoregressive decoding models include StructTrans, GVP, ProteinMPNN, and ESM-IF. The one-shot decoding model is represented by PiFold and the refinement decoding model is represented by ByProt. While

Metric Therefore, we propose to use "stability-based metric", which measure whether our structure-based design method can assign higher likelihood to the sequences with higher experimental stablity score. The score $\mathcal{R}$ is measured by:

$$\mathcal{R}(\theta, \mathcal{D}) = \rho_s\big(\mathcal{L}(S^{(i)}|X_{template}, \theta), \mathcal{G}^{(i)}\big) \tag{1}$$

where ρ_s is Spearman's correlation, θ is the design model, $\mathcal{L}$ is the pseudo-log-likelihood function and $\mathcal{D} = \big\{\mathcal{X}_{template}, \mathcal{S}^{(i)}, \mathcal{G}^{(i)}\big\}$ is the evaluation dataset, with $\mathcal{X}_{template}$ the template structure, $\mathcal{S}^{(i)}$ the i-th sequence and $\mathcal{G}^{(i)}$ the stability score corresponding to the i-th sequence. If the score $\mathcal{R}$ is high, the protein design method is likely to assign higher probability to the sequences with higher stability. Addressing the previously mentioned limitations of the perplexity metric, this dataset with multiple sequences can construct a more accurate sequence landscape that approximate the ground truth distribution. By the way, we apply the Spearman's correlation to calculate $\mathcal{R}$, which measure the correlation between score rankings instead of the direct relationship between two attributs, making the protein design methods comparable among them again. Note that the temperature is set to 1.0 for all the models.

Dataset We construct such datasets from two types of high-throughput data: "*De Novo* Design" [Rocklin et al., 2017] and "Mutagenesis" [Rocklin et al., 2017, Tsuboyama et al., 2023]. The statistics for these datasets are shown in Table.3, Table.4, and Table.5 while Figure.2a and Figure. 2b illustrate example datasets. The first category, *De Novo* Design data, refers to proteins modeled after specific structural templates. These proteins are designed based on these structure templates, and they will fold into corresponding structure once it can be folded. Even though there is a one-to-one relationship between structure and sequence in this data, structures stemming from the same topology show only subtle differences. For our curated dataset, we clustered these structural templates and replaced individual templates with the centroid of their respective clusters. Given that all structures within a cluster have a TMscore exceeding 0.5 with each other, it is reasonable to assume that sequences derived from these structures would have highly similar folds [Xu and Zhang, 2010]. In contrast, the second category, Mutagenesis data, is derived from various templates, including both natural proteins in the PDB and *De Novo* designed proteins with predicted structures. These datasets contain a significant amount of single-site and double-site mutation data related to the corresponding template, providing insight into which mutations stabilize the protein.[2] We further removed the 'insertion' and 'deletion' types of mutations, which alter the length of amino acid sequences, from the original dataset [Tsuboyama et al., 2023]. This resulted in 527K sequences with a stability score.



(a) (b)

Figure 2: (a) *De Novo* Design dataset with one structure template and four corresponding sequences along with stability scores. (b) Mutagenesis dataset with one structure template and four corresponding sequences along with stability scores.

4 Experiments

Baselines We evaluate StructTrans [Ingraham et al., 2019], GVP [Jing et al., 2020], ProteinMPNN [Dauparas et al., 2022], PiFold [Gao et al., 2023b], ByProt [Zheng et al., 2023], AF-Design[3] [Wang et al., 2022], ESM-Design [Verkuil et al., 2022], ESM-IF1 [Hsu et al., 2022] with our

ESM-Design is categorized under MCMC sampling models, AF-Design is a gradient descent-based model with $\lambda_{LM} = 0$.

[2]The stability-based metric evaluated on Mutagenesis dataset is similar to the experiment conducted by Ingraham et al. [2019]. However, while Ingraham et al. [2019] applied the Pearson correlation score.

[3]https://github.com/sokrypton/ColabDesign

Table 3: Dataset statistic of the *De Novo* Design data. We clustered the protein structures and selected the four largest clusters as datasets based on two design topologies: $EHEE$ and HHH. For instance, $EHEE_6$ denotes the 6th structural cluster, which is the largest cluster within the $EHEE$ topology. Sequences are considered stable if their experimental stability score is greater than or equal to 1. We are working on the structural clustering of *de novo* designed proteins to curate more datasets.

	$EHEE_3$	$EHEE_4$	$EHEE_5$	$EHEE_6$	HHH_{54}	HHH_{82}	HHH_{84}	HHH_{86}
# sequence	1743	1850	477	6873	669	632	1990	612
# stable sequence	110	120	31	511	203	186	621	213
portion	0.06	0.06	0.06	0.07	0.30	0.29	0.31	0.35

Table 4: Dataset statistic of the Mutagenesis data derived from Rocklin et al. [2017]. Sequences are considered stable if their experimental stability score is greater than or equal to 1

	$EEHEE_{37}$	$EEHEE_{1498}$	$EEHEE_{1702}$	$EEHEE_{1716}$	$EEHEE_{779}$
# sequence	775	775	775	775	775
# stable sequence	49	392	680	339	163
portion	0.063	0.506	0.877	0.437	0.21

	$HEEH_{223}$	$HEEH_{726}$	$HEEH_{872}$	HHH_{142}	HHH_{134}	HHH_{138}
# sequence	775	775	775	775	775	775
# stable sequence	453	39	438	623	720	754
portion	0.585	0.05	0.565	0.804	0.929	0.973

Table 5: Dataset statistic of the Mutagenesis data derived from dataset #3 in Tsuboyama et al. [2023]

Dataset	Description	# of total sequences	sequence group	# sequences groups	# of sequences
Original Dataset	All data for $\Delta\Delta G$ (WT < 4.75 kcal/mol)	607,839	single-site mutation	412 wild-types	448,788
			double-site mutation	496 pairs	159,051
Filtered Dataset	remove "indel" and "delete"	527,830	single-site mutation	372 wild-types	368,779
			double-site mutation	481 pairs	159,051

benchmark. Each model follows the default settings provided in their original papers or codebases. Encoder-decoder models were originally trained on the CATH4.2 train dataset for up to 100 epochs. ESM-IF1, on the other hand, was originally trained on the CATH4.3 data set[4]. ESM-Design and AF-Design models were trained on full UniRef data [Suzek et al., 2015]or complete PDB data [Jumper et al., 2021, Berman et al., 2000]. We chose to overlook potential data leakage issues for these models because they begin from random starting points and is inable to sample the exact ground truth sequence accurately, as demonstrated in further experiments. All experiments were conducted on Nvidia Quadro RTX8000.

4.1 Benchmarking on Refoldability-based Metrics

Settings For each structure in the test set, we randomly generated 100 sequences using protein design models and also from random models that sample from uniform and natural frequency distributions. The sampling temperature is set to 0.1 for all encoder-decoder models. However, due to the slow inference speed of AF-Design (2.4 GPU hours per sequence on average) and ESM-Design (9 GPU hours per sequence on average), we limited our generation to 5 sequences per structure for AF-Design and just 1 sequence per structure for ESM-Design[5]. We then predicted the structures of these generated sequences using both ESMFold and OmegaFold. AlphaFold2[6] is somehow time

[4]Since ESM-IF1 is trained on CATH4.3, we did not evaluate its refoldability on CATH4.2 to avoid potential data leakage. Currently, we cannot train ESM-IF on CATH4.2 because the training code has not been provided.

[5]would the results be statistical significant? Also this 1 vs. 5, would this cause bias? → Elaborate that, (1) this methods are too slow, we cannot afford training it, (2) with 1 sample generated, we can already have a coarse-grained observation on the performance, and see that their performance is largerly lagged behind other methods.

[6]We use the ColabFold [Mirdita et al., 2022] implementation with MMseqs MSA alignment [Steinegger and Söding, 2017, Mirdita et al., 2019].

costly to run, so we randomly feed one sequence per structure into AlphaFold2. Finally, we employ the TMalign toolkit [Zhang and Skolnick, 2005] to compute the Ref-TM score.

Table 6: Refoldability metric and recovery metric on the CATH dataset. We employ **bold** and underlining to highlight the best and suboptimal results on each metric. We use TM and pLDDT to represent Ref-TM and Ref-pLDDT.

Design method	ESMFold		OmegaFold		AlphaFold2		Recovery%
	TM	pLDDT	TM	pLDDT	TM	pLDDT	
Uniform	0.05	27.68	0.05	31.53	0.06	33.68	5.00
Natural frequencies	0.07	30.53	0.07	35.59	0.06	35.02	5.84
StructTrans	0.72	68.85	0.64	70.35	0.79	80.66	35.89
GVP	0.73	69.67	0.67	74.33	0.83	84.29	39.46
ProteinMPNN	**0.80**	**76.53**	**0.76**	**80.75**	**0.87**	**87.89**	41.44
PiFold	0.71	67.55	0.64	70.21	0.82	82.54	44.86
ByProt	0.73	72.12	0.70	77.58	0.85	87.26	**51.23**
AF-Design	0.53	61.37	0.53	72.04	0.52	75.29	15.95
ESM-Design	0.38	59.65	0.38	62.66	0.37	60.02	17.33
Wildtype	0.80	74.91	0.75	78.39	0.90	89.87	100

Refoldability Metric Analysis We report the refoldability and recovery metrics in Table. 6. We observe that ProteinMPNN stands out as the leading design method across the refoldability metrics, attaining 0.87 Ref-TM and 87.89 Ref-pLDDT with AlphaFold2 prediction. ByProt is slightly behind with a 0.85 Ref-TM and 87.26 Ref-pLDDT, followed by GVP. The sampling-based model ESM-Design and gradient-based model AF-Design are subpar in terms of both recovery and refoldability. We note that despite low recovery, the structure somewhat resembles the input structure.

Refoldability Metric on Ground Truth Sequence and Random Sequences To provide a context, we tested the refoldability of sequences generated by "non-learnable models" and wildtype sequences. (i) Sequences generated by random sampling have very low refoldability metric. Their Ref-TM is only around 0.05, suggesting a significant dissimilarity between the folded structure and the input structure. Additionally, the low pLDDT value implies poor sequence quality, indicating difficulty in proper folding. (ii) In contrast, for wildtype sequences, the AlphaFold2 model predicts a higher Ref-TM of 0.90 and Ref-pLDDT of 89.87. This suggests that we can trust AlphaFold's predictive accuracy for novel sequences.

Discord between Recovery and Refoldability Table. 6 also indicates that the recovery metric and refoldability metrics are not fully aligned. For example, (i) ProteinMPNN, which ranks 3rd in the recovery metric with a 41.44% recovery, lags behind ByProt's 51.23% recovery. However, ProteinMPNN achieves the highest scores on both Ref-TM and Ref-pLDDT metrics; (ii) PiFold, despite having the second-highest recovery at 44.86%, only ranks fourth in refoldability metrics, lagging behind ByProt, GVP, and ProteinMPNN.

Consistency between Structure Prediction Models We noticed that the ranking of different protein design methods remains consistent when we apply different structure prediction models. This further bolsters the credibility of the refoldability metrics. Moreover, if a protein design method excels under ESMFold, it will likely perform similarly well under OmegaFold and AlphaFold2.

Consistency between Ref-TM and Ref-pLDDT We find that the trends for Ref-TM and Ref-pLDDT computed based on the three structure prediction models are strikingly similar. When the Ref-pLDDT is higher, the Ref-TM is typically higher as well. Consequently, for structure-based protein design, we recommend to use structural prediction models as discriminators to pre-screen the generated sequences with Ref-pLDDT.

4.2 Benchmarking on Stability-based Metrics

Result Anaylsis on *De Novo* Design Data Table. 7 shows the stability metric on *De Novo* Design datasets. We observe that (i) AF-Design exhibits the highest correlation with stability scores, likely due to the utilization of AlphaFold2. However, sampling from the estimated distribution remains a challenge; (ii) Within the encoder-decoder family, ESMIF performs the best, followed by ByProt and ProteinMPNN; (iii) Surprisingly, ESM-Design does not perform as good as AF-Design model, and also falling short compared to other encoder-decoder methods.

Table 7: Stability metric on *De Novo* Design datasets.

Design method	$EHEE_3$	$EHEE_4$	$EHEE_5$	$EHEE_6$	HHH_{54}	HHH_{82}	HHH_{84}	HHH_{86}	mean
GVP	0.158	0.285	0.299	0.271	0.682	0.593	0.588	0.658	0.442
PiFold	0.158	0.287	0.269	0.267	0.688	0.607	0.556	0.641	0.434
ProteinMPNN	0.176	0.282	0.314	0.274	0.688	0.584	0.570	0.626	0.439
ESMIF	0.171	0.335	0.331	0.282	0.678	0.660	0.625	0.691	0.472
ByProt	0.191	0.297	0.296	0.270	0.688	0.631	0.571	0.626	0.446
AF-Design	**0.252**	**0.366**	**0.402**	**0.353**	**0.699**	**0.672**	**0.661**	**0.723**	**0.516**
ESM-Design	0.153	0.259	0.291	0.189	0.622	0.369	0.303	0.362	0.319

Result Anaylsis on Mutagenesis Data Table. 8 shows the stability metric on Mutagenesis datasets presented in [Rocklin et al., 2017]. (i) There is no single model that consistently performs well across all datasets. Overall, ESM-IF again achieves the highest mean correlation score of 0.433, and PiFold achieves the second with 0.413 correlation. (ii) The observation that PiFold performs well in density estimation on the mutational dataset and in recovery suggests that PiFold excels at modeling per-residue likelihood. (iii) The performance of AF-Design and ESM-Design is subpar. The possible reason is that structure prediction based models are not sensitive to point mutations [Pak et al., 2023].

Table 8: Stability metric on Mutagenesis datasets in [Rocklin et al., 2017].

Design method	$EEHEE_{37}$	$EEHEE_{1498}$	$EEHEE_{1702}$	$EEHEE_{1716}$	$EEHEE_{779}$	$HEEH_{223}$
GVP	0.481	0.318	0.247	0.413	0.526	0.340
PiFold	0.581	0.298	0.187	0.477	0.580	0.413
ProteinMPNN	0.597	0.382	0.136	0.384	0.595	0.324
ESMIF	**0.641**	0.382	0.236	**0.565**	**0.645**	**0.454**
ByProt	0.629	**0.414**	**0.320**	0.548	0.584	0.402
AF-Design	0.557	0.300	0.027	0.036	0.490	0.195
ESM-Design	0.240	0.115	-0.080	0.188	0.039	0.227

Design method	$HEEH_{726}$	$HEEH_{872}$	HHH_{142}	HHH_{134}	HHH_{138}	mean
GVP	0.102	0.248	0.502	0.253	0.295	0.339
PiFold	**0.239**	0.315	0.536	0.290	0.383	0.391
ProteinMPNN	-0.055	0.205	0.431	0.256	0.326	0.326
ESMIF	0.216	0.335	**0.573**	0.318	**0.398**	**0.433**
ByProt	0.238	**0.338**	0.511	0.289	0.360	0.421
AF-Design	0.214	-0.148	0.453	**0.351**	0.314	0.254
ESM-Design	0.062	0.013	0.004	-0.050	-0.050	0.064

Result Anaylsis on Mega-Scale Mutagenesis Data Table 9 presents the stability metric applied to mega-scale Mutagenesis datasets, which includes 527,830 sequences. This dataset is significantly larger than the one comprising 8,525 sequences used in the previous table. We have divided the dataset into two parts: mutations on 215 natural proteins and mutations on 156 *de novo* designed proteins. This division allows us to examine whether the models perform differently on these groups. The correlation scores were first calculated for each sequence group relative to its corresponding wild-type protein, and then these scores were averaged. Our observations are as follows: (i) ESMIF consistently achieves the highest correlation scores across both *de novo* proteins and natural proteins, followed by ProteinMPNN, ByProt, and PiFold; (ii) Encoder-decoder based models show lower correlation scores on *de novo* sequence groups, while structure-prediction based models attain higher scores on natural sequence groups; (iii) The performance of AF-Design and ESM-Design remains subpar in this larger dataset. Notably, ESM-Design performs poorly on natural proteins, exhibiting

both positive and negative Spearman's correlation, which results in an average correlation score near zero.

Table 9: Stability metric applied on mega-scale experimental Mutagenesis datasets [Tsuboyama et al., 2023]. The columns display the average stability scores for *de novo* designed proteins, natural proteins in the PDB, and across all 372 sequence groups.

Design method	*De Novo*	Natural	All
GVP	0.390	0.494	0.450
PiFold	0.448	0.556	0.511
ProteinMPNN	0.428	<u>0.605</u>	0.531
ESMIF	**0.500**	**0.629**	**0.575**
ByProt	<u>0.468</u>	0.586	<u>0.536</u>
AF-Design	0.354	0.292	0.318
ESM-Design	0.127	0.0004	0.053

4.3 Takeaways

- Our findings indicate differences between the recovery metric and the refoldability metrics introduced in our benchmark. Notably, a model with high recovery doesn't necessarily guarantee good refoldability.
- Results suggest that encoder-decoder methods generally outperform structure-prediction based methods in terms of refoldability, recovery, and stability metrics. However, the structure-prediction based methods show potential in accurate sequence density estimation, which may lead to the generation of superior sequences.
- Among encoder-decoder methods, ByProt, ProteinMPNN, and ESM-IF show strong performance on our benchmark. We observe that PiFold excels in recovery metrics, but its performance in refoldability and stability metrics is less impressive. This may potentially be related to the conditional independence assumption applied during sampling and density estimation.
- Among structure-prediction based methods, AF-Design method has advantages over the ESM-Design method in various metrics, also including inference efficiency.

5 Conclusion and Future Work

Conclusion To better evaluate structure-based protein design models, we propose the refoldability-based metric and stability-based metric. We curate datasets corresponding to these metrics, and conduct experiments on this **PDB-Struct** benchmark. By examining the benchmark results, we detail the strengths and weaknesses of each type of protein design model, offering insights to protein researchers in their model selection. This paves the way for a fair and comprehensive evaluation of protein design methods in the future.

Future work We are continuously collecting additional *De Novo* Design and Mutagenesis datasets to enhance our benchmark, and we are evaluating newly released protein-design methods such as KW-design [Gao et al., 2023a] and GRADE-IF [Yi et al., 2023]. Furthermore, we are conducting extensive experiments to demonstrate the superiority of refoldability metrics over the recovery metric. Discussions regarding the efficiency and reliability of the **PDB-Struct** benchmark evaluations are ongoing, and we intend to address these in a future version of this work. Concurrently, we discovered another project for benchmarking protein design methods, ProteinInvBench [Gao et al., 2023c], which has been accepted into the NeurIPS 2023 Datasets and Benchmarks Track. Inspired by their approach, we are considering the addition of a diversity metric to our benchmark.

Acknowledgement

We would like to thank to Zuobai Zhang, Yangtian Zhang for their constructive feedback on the manuscript. This project is supported by Twitter, Intel, the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, the Canada CIFAR AI Chair Program, Samsung Electronics Co., Ltd., Amazon Faculty Research Award, Tencent AI Lab Rhino-Bird Gift Fund, a NRC Collaborative R&D Project (AI4D-CORE-06) as well as the IVADO Fundamental Research Project grant PRF-2019-3583139727.

References

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, pages 2023–09, 2023.

Namrata Anand, Raphael Eguchi, Irimpan I Mathews, Carla P Perez, Alexander Derry, Russ B Altman, and Po-Ssu Huang. Protein sequence design with a learned potential. *Nature communications*, 13 (1):746, 2022.

Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. How cryo-em is revolutionizing structural biology. *Trends in biochemical sciences*, 40(1):49–57, 2015.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

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

Fred E Cohen and Jeffery W Kelly. Therapeutic approaches to protein-misfolding diseases. *Nature*, 426(6968):905–909, 2003.

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.

Zhangyang Gao, Cheng Tan, and Stan Z Li. Knowledge-design: Pushing the limit of protein deign via knowledge refinement. *arXiv preprint arXiv:2305.15151*, 2023a.

Zhangyang Gao, Cheng Tan, and Stan Z. Li. Pifold: Toward effective and efficient protein inverse folding. In *International Conference on Learning Representations*, 2023b. URL `https://openreview.net/forum?id=oMsN9TYwJ0j`.

Zhangyang Gao, Cheng Tan, Yijie Zhang, Xingran Chen, Lirong Wu, and Stan Z Li. Proteininvbench: Benchmarking protein inverse folding on diverse tasks, models, and metrics. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023c.

Brian Hie, Salvatore Candido, Zeming Lin, Ori Kabeli, Roshan Rao, Nikita Smetanin, Tom Sercu, and Alexander Rives. A high-level programming language for generative protein design. *bioRxiv*, pages 2022–12, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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. *ICML*, 2022. doi: 10.1101/2022.04.10.487779. URL `https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779`.

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

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, et al. Illuminating protein space with a programmable generative model. *BioRxiv*, pages 2022–12, 2022.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

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.

Marcus Frederick Charles Ladd, Rex Alfred Palmer, and Rex Alfred Palmer. *Structure determination by X-ray crystallography*, volume 233. Springer, 1977.

Zhixiu Li, Yuedong Yang, Eshel Faraggi, Jian Zhan, and Yaoqi Zhou. Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins: Structure, Function, and Bioinformatics*, 82(10): 2565–2573, 2014.

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.

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

Sidney Lyayuga Lisanza, Jacob Merle Gershon, Sam Wayne Kenmore Tipps, Lucas Arnoldt, Samuel Hendel, Jeremiah Nelson Sims, Xinting Li, and David Baker. Joint generation of protein sequence and structure with rosettafold sequence space diffusion. *bioRxiv*, pages 2023–05, 2023.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. 2022. URL `https://openreview.net/forum?id=jSorGn2Tjg`.

Weian Mao, Muzhi Zhu, Hao Chen, and Chunhua Shen. Modeling protein structure using geometric vector field networks. *bioRxiv*, pages 2023–05, 2023.

Milot Mirdita, Martin Steinegger, and Johannes Söding. Mmseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, 35(16):2856–2858, 2019.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Sadao Mori and Howard G Barth. *Size exclusion chromatography*. Springer Science & Business Media, 1999.

James O'Connell, Zhixiu Li, Jack Hanson, Rhys Heffernan, James Lyons, Kuldip Paliwal, Abdollah Dehzangi, Yuedong Yang, and Yaoqi Zhou. Spin2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Structure, Function, and Bioinformatics*, 86(6): 629–633, 2018.

Christine A Orengo, Alex D Michie, Susan Jones, David T Jones, Mark B Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Marina A Pak, Karina A Markhieva, Mariia S Novikova, Dmitry S Petrov, Ilya S Vorobyev, Ekaterina S Maksimova, Fyodor A Kondrashov, and Dmitry N Ivankov. Using alphafold to predict the impact of single mutations on protein stability and function. *Plos one*, 18(3):e0282689, 2023.

Yifei Qi and John ZH Zhang. Densecpd: improving the accuracy of neural-network-based computational protein sequence design with densenet. *Journal of chemical information and modeling*, 60 (3):1245–1252, 2020.

Bao-He Qu, Elizabeth Strickland, and Philip J Thomas. Cystic fibrosis: a disease of altered protein folding. *Journal of Bioenergetics and Biomembranes*, 29:483–490, 1997.

Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357 (6347):168–175, 2017.

Chence Shi, Chuanrui Wang, Jiarui Lu, Bozitao Zhong, and Jian Tang. Protein sequence and structure co-design with equivariant translation. *arXiv preprint arXiv:2210.08761*, 2022.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Cheng Tan, Zhangyang Gao, Jun Xia, Bozhen Hu, and Stan Z Li. Generative de novo protein design with global context. *arXiv preprint arXiv:2204.10673*, 2022.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi S. Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6TxBxqNME1Y`.

Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J Weinstein, Niall M Mangan, Sergey Ovchinnikov, and Gabriel J Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973): 434–444, 2023.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

Robert Verkuil, Ori Kabeli, Yilun Du, Basile IM Wicky, Lukas F Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. *bioRxiv*, pages 2022–12, 2022.

Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

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. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *BioRxiv*, pages 2022–12, 2022.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pages 2022–07, 2022.

Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, 2020.

Andy Hsien-Wei Yeh, Christoffer Norn, Yakov Kipnis, Doug Tischer, Samuel J Pellock, Declan Evans, Pengchen Ma, Gyu Rie Lee, Jason Z Zhang, Ivan Anishchenko, et al. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023.

Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yu Guang Wang. Graph denoising diffusion for inverse protein folding. *arXiv preprint arXiv:2306.16819*, 2023.

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133*, 2022.

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. *bioRxiv*, pages 2023–02, 2023.