Benchmark of Diffusion and Flow Matching Models for Unconditional Protein Structure Design

Wen-ran Li[*1] Xavier F. Cadet[*2] Cédric Damour[3] Yu Li[4] Alexandre G. de Brevern[1] Alain Miranville[5] Frederic Cadet[1 6]

Abstract

With the widespread application of deep neural networks, generative models, especially diffusion models and flow matching models, for protein design have experienced explosive growth. However, there remains a lack of comprehensive evaluation frameworks to systematically assess the performance of these models. This study addresses this gap by focusing on the task of designing unconditional protein structures, benchmarking seven state-of-the-art (SOTA) models in four distinct dimensions: structural validity, diversity, novelty, and computational efficiency. This work provides standardized metrics and baseline benchmarks to guide future research and innovation in protein design.

1. Introduction

Diffusion models (Ho et al., 2020) and flow matching models (Lipman et al., 2024) share a similar architecture, and their application to protein generation is gaining attention. However, which kinds of models are better for the generation of protein structure have not been systematically studied. This paper examines the performance of 6 diffusion models and 1 flow matching model across unconditional generation and structure prediction. This work provides valuable guidance to help researchers select the most appropriate models for specific tasks and requirements.

Contributions: We investigate the theoretical foundations and empirical performance of recent diffusion-based and flow-based generative models. We provide a comparison of state-of-the-art approaches and propose insights for combining their strengths; in this work:

- We present the foundational theory of diffusion and flow matching models and introduce 6 state-of-the-art (SOTA) diffusion-based models and 1 SOTA flow-based model.
- We propose 4 benchmarking dimensions to evaluate the models and conduct a comprehensive evaluation and ranking based on these criteria.
- We analyze the results of the evaluation to highlight the strengths of the flow matching model and explore a strategy to integrate diffusion and flow matching approaches for improved performance.

2. Model introduction

There are 7 models discussed in this paper, 6 of them are diffusion models, and 1 is flow matching. The two types of models are very similar. They both involve the process of adding noise and removing noise. In some cases, they are even equivalent. The forward process of the diffusion model which gradually destroys data over time can be described by the following stochastic differential equations (SDE):

$$d\mathbf{z}_t = f_t \mathbf{z}_t dt + g_t d\omega$$

where $d\omega$ is a Brownian motion, f_t and g_t decide the noise schedule. The generative process is given by

$$d\mathbf{z}_t = (f_t \mathbf{z}_t - \frac{1+\eta_t^2}{2} g_t^2 \nabla \log p_t(\mathbf{z}_t))dt + \eta_t g_t d\omega,$$

where $\nabla \log p_t$ is the score of the forward process. η_t controls the amount of stochasticity at inference time. The interpolation between x and ϵ in flow matching can be described by the following ordinary differential equation (ODE):

$$d\mathbf{z}_t = \mathbf{u}_t dt,$$

[*]Equal contribution [1]University of Paris City and University of Reunion, INSERM, BIGR, DSIMB Bioinformatics Team, F-75015 Paris, France [2]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755, USA (Work done whilst at Department of Computing, Imperial College London) [3]EnergyLab, EA 4079, Faculty of Sciences and Technology, University of Reunion, France [4]School of Information Science and Technology and Beijing Institute of Artificial Intelligence, Beijing, China [5]Laboratoire de Mathématiques Appliquées du Havre (LMAH), University Le Havre Normandie, Le Havre, France [6]PEACCEL, Artificial Intelligence Department, AI for Biologics, 75013, Paris, France. Correspondence to: Frederic Cadet <frederic.cadet.run@gmail.com>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

assuming the interpolant is $\mathbf{z}_t = \alpha_t x + \sigma\epsilon$, then $\mathbf{u}_t = \dot{\alpha_t} x + \dot{\sigma_t}\epsilon$, where $\dot{\alpha_t}$ denotes the derivative of α_t with respect to t. The generative process can be generalized to SDE:

$$d\mathbf{z}_t = (\mathbf{u}_t - \frac{1}{2}\epsilon_t^2 \nabla \log p_t(\mathbf{z}_t))dt + \epsilon_t dt,$$

where ϵ_t modulates the noise level at inference time.

We can deduce the equivalence by deriving on set of hyper-parameters for the other. From diffusion to flow matching:

$$\begin{aligned} \alpha_t &= \exp(\int f_s ds), \quad \epsilon_t = \eta_t g_t, \\ \sigma_t &= (\int_0^t g_s^2 \exp(-2\int_0^s f_u du))^{\frac{1}{2}}. \end{aligned} \tag{1}$$

From flow matching to diffusion:

$$\begin{aligned} f_t &= \partial_t \log(\alpha_t), \quad g_t^2 = 2\alpha_t\sigma_t\partial_t(\frac{\sigma_t}{\alpha_t}), \\ \eta_t &= \epsilon_t/(2\alpha_t\sigma_t\partial_t(\frac{\sigma_t}{\alpha_t}))^{\frac{1}{2}}. \end{aligned} \tag{2}$$

In summary, apart from training consideration and sampler selection, diffusion and flow matching exhibit no fundamental differences. However, there are two new model specifications (Gao et al., 2024) that flow matching brings to the field:

- **Network output:** Flow matching proposes a vector field parameterization of the network output that is different from the ones used in diffusion literature. The network output can make a difference when higher-order samplers are used. It may also affect the training dynamics.
- **Sampling noise schedule:** Flow matching leverages a simple sampling noise schedule.

Here is a general introduction to the 7 models:

RFDiffusion (Watson et al., 2023) takes the classical RoseTTAFold (Baek et al., 2021) as the reverse diffusion denoiser step to design the protein structure. **FrameDiff** (Yim et al., 2023b) uses the Invariant Point Attention (IPA) (Jumper et al., 2021) module for keeping $SE(3)$ invariance.

Genie (Lin & AlQuraishi, 2023) uses IPA to construct a mechanism for computationally efficient $SE(3)$ equivariant diffusion denoiser to generate protein backbones. **Genie2** (Lin et al., 2024) adds motif scaffolding capabilities via a novel multi-motif framework that designs co-occurring motifs with unspecified inter-motif positions and orientations.

ProtDiff-SMCDiff (Trippe et al., 2023) is a diffusion model for motif-scaffolding using sequential Monte Carlo (SMC). **FoldingDiff** (Wu et al., 2024) predicts the angles between the residues, but not the absolute location of the residues.

For flow matching model, **FrameFlow** (Yim et al., 2023a) adapts FrameDiff to the flow-matching generative modeling paradigm, keeps model $SE(3)$ equivariance and achieves twice the design capability compared to FrameDiff.

3. Unconditional structure generation benchmarking

Unconditional structure generation is defined as a structure generated by a pre-trained model without inputting any sequence or structure. Using these 7 pre-trained models, 100 proteins were unconditionally generated with lengths of 50, 100, 200, 300, 400, and 500 residues, respectively.

A multimetric evaluation is created to assess the models' performance on unconditional generation, encompassing 4 key dimensions: designability, novelty, diversity, and efficiency. The metrics for evaluating them are listed in Table 1. In this section, the evaluation and the corresponding results are presented in different subsections, followed by a comprehensive assessment of their performance.

3.1. Designability

The **designability** to design a structure reflects the ability to identify an amino acid sequence that can fold into the designed backbone structure. Following the protocol of (Trippe et al., 2023), 100 structures were generated from the models to be tested. For each generated structure, we input it to ProteinMPNN (Dauparas et al., 2022) for inverse folding 10 sequences. The total of 1,000 generated sequences was input into ESMFold (Manfredi et al., 2025) to refold structures, see Fig. 2 in appendix. There are two methods to evaluate the predictability of the different models:

- Calculating the **scTM** between the generated structure and the refolded structure for evaluation.
- Calculating scRMSD (similar to scTM but using RMSD) between generated and refolded structures for evaluation.

3.2. Novelty

Novelty measures the rate at which the generated samples differ significantly from the reference set. Taking into account the novelty raised by (Yim et al., 2023b), Foldseek (van Kempen et al., 2022) used the maximum TM score between the generated protein set and the PDB dataset `https://www.rcsb.org/`, with the value recorded as **pdbTM** to evaluate novelty.

3.3. Diversity

A large **diversity** (Yim et al., 2023b; Zheng et al., 2024) means that dissimilar sequences and a large coverage of the

Table 1. List of metrics used in this paper. The range related to TM-score are $[0, 1]$, scRMSD ≥ 0, Max-Cluster $\in [0, 1]$, runtime. ↑ means the higher the better, ↓ means the lower the better.

Dimension	Metric	Definition	Range	Ideal Direction
Designability	scTM	TM-score between generated and refolded structures	[0,1]	↑
	scRMSD	C_α RMSD between generated and refolded structures	≥ 0	↓
Novelty	pdbTM	Highest TM-score between generation structures and pdb dataset.	[0,1]	↓
Diversity	Pairwise TM	Max TM-score across all generated structure pairs	[0,1]	↓
	Max-Cluster	Proportion of clusters (TM-score threshold = 0.5)	[0,1]	↑
Efficiency	Run time	The time spend for generating 100 proteins	≥ 0	↓

sequence space can fold into a given structure. Diversity is measured using two methods, both based on Pairwise TM-score, which means the TM-score(p_1, p_2) for any p_1, p_2 in the generated protein set:

- The **maximum pairwise TM-score** is used as a diversity metric. In the topological space, the lower the TM-score, the longer the distance between two proteins, the more dissimilar they are, the higher the diversity.
- Clustering the generated protein backbones is a way to assess the similarities between protein backbones. **Max-Cluster** works by grouping similar protein structures into clusters based on their structural similarity, which is measured using the TM-score. In this paper, the clustering threshold is set as TM-score=0.5.

3.4. Efficiency

We run models on Nvidia L4 GPU×4, CUDA version 12.2, to generate 100 proteins per length, record the time, and list them in Table 2 to show their efficiency.

3.5. Structural Properties

In addition to the above 4 dimensions, some structural properties can also be extracted from the generated structures. For example, **secondary structure elements** (SSE) representing the structural organization, and **radius of gyration** measuring the size or compactness. Since there are no oxygen atoms O in the PDB file generated from FoldingDiff, Genie, Genie2 and ProtDiff-SMCDiff, it is impossible to accurately calculate the SSE. This paper just analyses the structural properties of FrameDiff, FrameFlow and RFDiffusion. DSSP (Touw et al., 2015) is a database of secondary structure assignments for all protein entries in the Protein Data Bank. Leveraging the DSSP (Sekihara et al., 2016) of Python package Biotite (Kunzmann et al., 2023), we compile statistics for the proportion of α-helix, β-strand, random coil and turn. In addition, the average radius of gyration is calculated according to the formula in (Xiang et al., 2013). The result can be seen in Figure 1. These properties exhibit natural biological variations without absolute optimal ranges, and therefore were not incorporated into holistic performance assessment.

4. Results

Table 2 lists the results, and Fig. 4 in the appendix allows visualization. All the results are discussed in this section, and then we give them a comprehensive evaluation and propose directions for improving the model for protein generation.

For the 2 metrics and 6 different lengths, the best results are highlighted in Table 2 in bold. Regarding **designability**, 6 numbers are highlighted for RFDiffusion, 4 numbers for Frameflow, 1 for FrameDiff and 1 for Genie. RFdiffusion shows the best designability among these 7 models. It can also be observed that if the designability is good with respect to scTM, it is also good for scRMSD. In fact, except for these 2 metrics, there are some other metrics to show the designability, such as PAE, pTM, and pLDDT, for comparison between the generated and refolded protein structures. Using the results from FrameFlow as an example, Fig. 3 in appendix shows the relation between the 5 metrics: scTM, scRMSD, pTM, pLDDT and PAE. It can be seen that the five metrics that describe designability show strong positive or negative correlations with each other.

In generating proteins of 6 different lengths, ProtDiff-SMCDiff achieved the best **novelty** results in five of these cases, which can be attributed to the unique $E(3)$-invariant framework inherent to the ProtDiff-SMCDiff model. While Table 2 lists the average pdbTM values, we can further examine the distribution of pdbTM values in Fig. 4. The distribution of pdbTM for ProtDiff-SMCDiff is far lower than that of other models. Furthermore, a comparative analysis could be extended by substituting the reference dataset PDB with alternative datasets such as AlphaFold DB (Varadi et al., 2024) to validate the robustness of these findings.

As for **diversity**, Foldingdiff shows good performance when generating 50 amino acids and 100 amino acids. However, the test dataset in the original FoldingDiff paper has an upper limitation of just 128, i.e., this model can only generate proteins of less than 128 amino acids. See Fig. 4 (b) for a

Table 2. Result of unconditional generation. scTM (↑) and scRMSD (↓) for designability evaluation, pdbTM (↑) for novelty evaluation, Pairwise TM-score (↓) and Max cluster (↑) for diversity evaluation and runtime (↓) for efficiency evaluation. Performance score (↑) calculated by TOPSIS for rank.

	Model	Designability		Novelty	Diversity		Efficiency	Performance score	Rank
		scTM	scRMSD	pdbTM	Pairwise TM	Max Clust.	Runtime		
50	FrameDiff	0.781	0.568	0.715	0.870	0.490	1h-25m-41.52s	0.540	5
	FrameFlow	0.807	0.476	0.727	0.925	0.320	6m-27.73s	0.587	3
	Genie	**0.841**	6.079	0.699	0.848	0.590	3h-20m-26s	0.319	7
	Genie2	0.400	5.602	0.743	0.864	0.370	2h-1m-55s	0.482	6
	FoldingDiff	0.289	9.874	0.599	0.933	**0.820**	3m-30.26s	0.611	2
	RFDiffusion	0.726	**0.270**	0.639	0.981	0.120	53m-45s	0.543	4
	ProtDiff-SMCDiff	0.348	8.289	**0.500**	**0.640**	0.930	**3m-28s**	**0.640**	**1**
100	FrameDiff	0.799	0.786	0.683	0.932	0.460	1h-36m-49.11s	0.628	3
	FrameFlow	0.862	0.605	0.678	0.729	0.400	5m-42.14s	**0.737**	**1**
	Genie	0.836	10.899	0.595	0.749	0.830	3h-26m-25.11s	0.372	7
	Genie2	0.354	9.354	0.644	0.631	0.560	1h-2m-39s	0.502	6
	FoldingDiff	0.298	14.583	**0.563**	**0.591**	**0.990**	**2m-26s**	0.566	4
	RFDiffusion	**0.956**	**0.487**	0.709	0.781	0.301	59m-45s	0.659	2
	ProtDiff-SMCDiff	0.302	9.707	0.640	0.957	0.130	15m-31s	0.506	5
200	FrameDiff	0.694	**0.666**	0.676	0.743	0.390	3h-29m-59s	0.521	4
	FrameFlow	**0.989**	0.745	0.532	0.677	0.610	**11m-42s**	**0.818**	**1**
	Genie	0.280	16.265	0.586	0.606	0.670	3h-19m-18.08s	0.413	5
	Genie2	0.267	16.535	0.580	0.567	0.780	6h-8m-53.81s	0.154	6
	RFDiffusion	0.952	0.723	0.523	0.736	0.647	2h-0m-17s	0.731	2
	ProtDiff-SMCDiff	0.314	18.719	**0.391**	**0.329**	**0.950**	60m-29s	0.624	3
300	FrameDiff	0.682	1.670	0.642	0.746	0.480	23h-58m-1s	0.291	5
	FrameFlow	**0.980**	1.224	0.653	0.647	0.570	**20m-31s**	0.804	2
	Genie2	0.277	20.639	0.579	**0.350**	0.880	7h-9m-42.13s	0.563	4
	RFDiffusion	0.884	**0.722**	0.607	0.627	0.670	4h-25m-45s	**0.821**	**1**
	ProtDiff-SMCDiff	0.294	21.137	**0.350**	0.643	**0.900**	2h-16m-21s	0.679	3
400	FrameDiff	0.500	8.795	0.606	0.597	0.550	24h-33m	0.294	5
	FrameFlow	0.980	3.119	0.622	0.600	0.560	**2h-8m-35s**	**0.772**	**1**
	Genie2	0.233	37.086	0.571	**0.469**	0.870	9h-13m-45.32s	0.514	4
	RFDiffusion	**0.992**	**2.2106**	0.599	0.664	0.760	11h-48m-53s	0.653	2
	ProtDiff-SMCDiff	0.230	41.613	**0.363**	0.698	**0.990**	4h-59m-57s	0.621	3
500	FrameDiff	0.448	14.578	0.586	0.589	0.590	37h-6m-11s	0.289	5
	FrameFlow	**0.946**	**2.696**	0.627	0.595	0.540	**2h-20m-52s**	**0.789**	**1**
	Genie2	0.205	52.059	0.566	**0.456**	**0.920**	13h-24m-35s	0.505	4
	RFDiffusion	0.722	4.028	0.599	0.593	0.870	20h-4m-18s	0.593	2
	ProtDiff-SMCDiff	0.150	63.080	**0.264**	0.642	0.610	11h-20m-41s	0.536	3

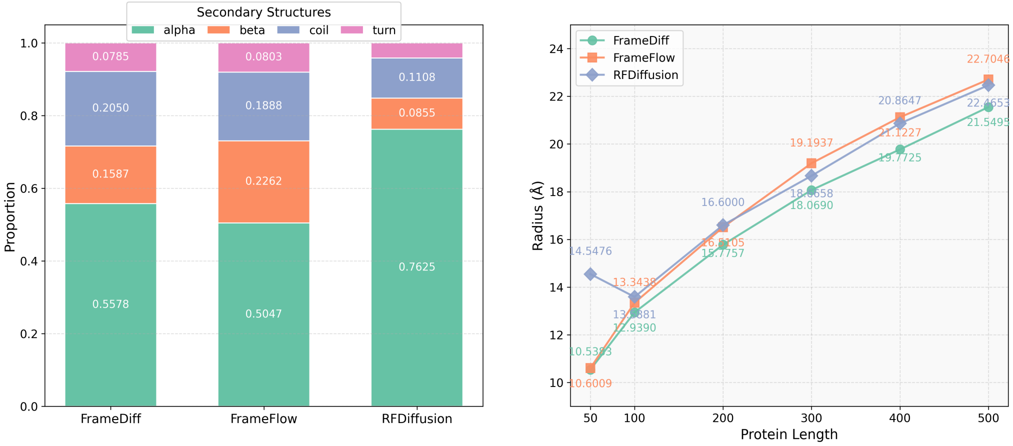

Figure 1. Results of secondary structure elements and radius of gyration. **Left:** Mean secondary structure composition (α, β, coil, turn) in 100 generated proteins (each protein has 100 amino acids) per method. **Right:** Length-dependent radius of gyration patterns.

time chart showing changes in metrics over length; the line changes of the 2 metrics have opposite trends.

For **efficiency**, in Fig. 4(c), the variation in runtime with protein length is illustrated, with an exponential function fitted to the data of each model. The form of the fitted exponential function is provided, and the corresponding coefficients a, b, and c for each model are listed in Table 3 in appendix. Since FrameFlow has the smallest exponential function and the most gradual increase, it is the most efficient model. It is worth noting that FoldingDiff is also a highly efficient model although it is not possible to obtain an exponential fitting for this model.

We employed the TOPSIS (Behzadian et al., 2012) method to calculate comprehensive performance scores and establish model rankings. Four key dimensions - designability, novelty, diversity, and efficiency - were assigned equal weights of 1/4 each. Within the designability dimension, the scTM and scRMSD metrics received sub-weights of 0.125 each, while in the diversity dimension, pairwise TM-score and Max-cluster were similarly weighted at 0.125 each. The calculation results (detailed in Table 2) show that FrameFlow achieved the highest comprehensive ranking (Rank 1) under this evaluation framework.

As outlined in Section 2, the flow matching model based on ODEs features a simplified sampling process. This fundamental characteristic explains why FrameFlow derived from FrameDiff achieves a 5x faster sampling speed while maintaining comparable performance. Future research directions could explore the implementation of denoising implicit diffusion models (Song et al., 2020) for protein generation tasks. Notably, DDIM shares the same efficient sampling paradigm as flow matching methods, making it potentially more computationally efficient than traditional diffusion approaches while preserving generation quality.

Acknowledgements

W. Li is supported by a PhD grant from the Region Reunion and European Union (FEDER-FSE 2021/2027) file number 2023062, tiers 345879. X. F. Cadet was supported by the UKRI CDT in AI for Healthcare http://ai4health.io, (Grant No. P/S023283/1), UK. PEACCEL was supported through a research program partially co-funded by the European Union (UE) and Region Reunion (FEDER). We gratefully acknowledge Google Cloud for providing supportive access to their computational clusters.

Impact Statement

This work introduces an open, four-dimensional benchmark that transparently ranks seven diffusion- and flow-matching models for de novo protein design, promoting reproducibility and faster progress in drug discovery, enzyme engineering and sustainable biocatalysis. By reducing trial-and-error in wet labs, it can lower R&D costs and environmental footprints. However, stronger protein generators also lower the barrier to designing harmful or immune-evasive biomolecules. Benchmark datasets remain biased toward well-studied organisms, which could perpetuate global health inequities, and large-scale model training incurs significant energy use. We therefore release only evaluation code and aggregated metrics under a licence discouraging unsafe use, recommend integrating bio-risk screening into future benchmarks, encourage expansion to under-represented taxa, and report compute footprints to foster greener research.

References

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch,

L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. 373(6557):871–876, 2021. ISSN 0036-8075, 1095-9203.

Behzadian, M., Otaghsara, S. K., Yazdani, M., and Ignatius, J. A state-of-the-art survey of topsis applications. *Expert Systems with applications*, 39(17):13051–13069, 2012.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning based protein sequence design using proteinmpnn. *bioRxiv*, 2022.

Gao, R., Hoogeboom, E., Heek, J., Bortoli, V. D., Murphy, K. P., and Salimans, T. Diffusion meets flow matching: Two sides of the same coin. 2024. URL https://diffusionflow.github.io/.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. 596(7873):583–589, 2021. ISSN 0028-0836, 1476-4687.

Kunzmann, P., Müller, T. D., Greil, M., Krumbach, J. H., Anter, J. M., Bauer, D., Islam, F., and Hamacher, K. Biotite: new tools for a versatile python bioinformatics library. *BMC bioinformatics*, 24(1):236, 2023.

Lin, Y. and AlQuraishi, M. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds, June 2023.

Lin, Y., Lee, M., Zhang, Z., and AlQuraishi, M. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie 2, May 2024.

Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R. T., Lopez-Paz, D., Ben-Hamu, H., and Gat, I. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Manfredi, M., Vazzana, G., Savojardo, C., Martelli, P. L., and Casadio, R. Alphafold2 and esmfold: A large-scale pairwise model comparison of human enzymes upon pfam functional annotation. *Computational and Structural Biotechnology Journal*, 27:461–466, 2025.

Sekihara, K., Kawabata, Y., Ushio, S., Sumiya, S., Kawabata, S., Adachi, Y., and Nagarajan, S. S. Dual signal subspace projection (dssp): a novel algorithm for removing large interference in biomagnetic measurements. *Journal of Neural Engineering*, 13(3):036007, 2016.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Touw, W. G., Baakman, C., Black, J., Te Beek, T. A., Krieger, E., Joosten, R. P., and Vriend, G. A series of pdb-related databanks for everyday needs. *Nucleic acids research*, 43(D1):D364–D368, 2015.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem, March 2023.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Gilchrist, C. L., Söding, J., and Steinegger, M. Foldseek: fast and accurate protein structure search. *Biorxiv*, pp. 2022–02, 2022.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, August 2023. ISSN 0028-0836, 1476-4687.

Wu, K. E., Yang, K. K., van den Berg, R., Alamdari, S., Zou, J. Y., Lu, A. X., and Amini, A. P. Protein structure generation via folding diffusion. *Nature communications*, 15(1):1059, 2024.

Xiang, S., Gapsys, V., Kim, H.-Y., Bessonov, S., Hsiao, H.-H., Möhlmann, S., Klaukien, V., Ficner, R., Becker, S., Urlaub, H., et al. Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. *Structure*, 21(12): 2162–2174, 2013.

Yim, J., Campbell, A., Foong, A. Y. K., Gastegger, M., Jiménez-Luna, J., Lewis, S., Satorras, V. G., Veeling, B. S., Barzilay, R., Jaakkola, T., and Noé, F. Fast protein backbone generation with se(3) flow matching, October 2023a.

Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023b.

Zheng, Z., Zhang, B., Zhong, B., Liu, K., Li, Z., Zhu, J., Yu, J., Wei, T., and Chen, H.-F. Scaffold-lab: Critical evaluation and ranking of protein backbone generation methods in a unified framework. *bioRxiv*, pp. 2024–02, 2024.

A. Outline of Appendix

This appendix includes:

- **Fig. 2**: Workflow for designability evaluation.
- **Fig. 3:** Scatter plot to reveal the relation among different metrics for designability.
- **Fig. 4:** Visualization of the results, Violin plots for designability and novelty, line charts for diversity and efficiency.
- **Table 3:** Efficiency: Coefficients of exponential fitting for the runtime with respect to length.
- List of abbreviations.

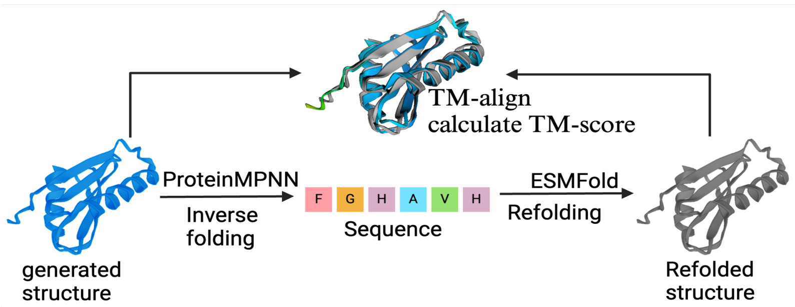

Figure 2. Designability workflow: from the generated structure, inverse folding is done and then refolded, both are compared using TM-score.

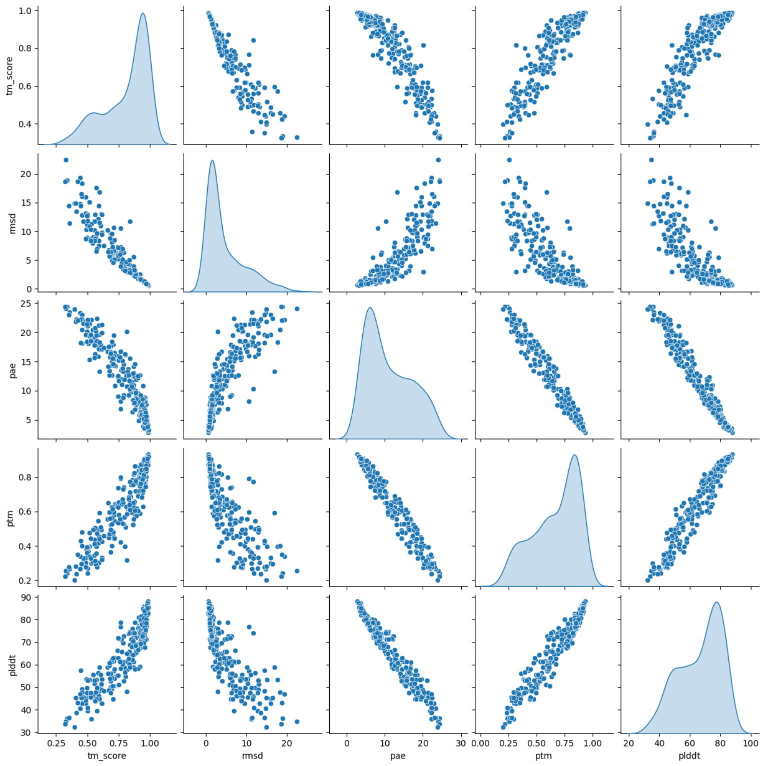

Figure 3. The relationship between scTM, scRMSD, plDDT, pTM, and PAE. The data come from FrameFlow for lengths 100, 300, and 500, respectively. They have high relevance in expressing the designability.

B. More details about the results

Table 3. **Efficiency.** Coefficients for the exponential fitting.

Model	a	b	c
FrameDiff	0.994	0.741	-1.09
FrameFlow	0.177	2.010	-0.241
Genie	37.1	-0.000435	-36.1
Genie2	1.41	0.510	-1.38
RFDiffusion	0.0521	3.02	-0.0505
ProtDiff-SMCDiff	0.021	3.88	-0.02

List of Abbreviations

$SE(3)$	special euclidean 3D group
DDPM	Denoising Diffusion Probabilistic Models
ODE	Ordinary Differential Equation
PAE	Predicted Alignment Error
pLDDT	Predicted Local Distance Different Test
RFDiffusion	RoseTTAFold Diffusion
RMSD	Root Mean Square Deviation
scTM	Self consistency Template Modeling score
SDE	Stochastic Differential Equation
SGM	Score-based Generative Models
SOTA	State of the art

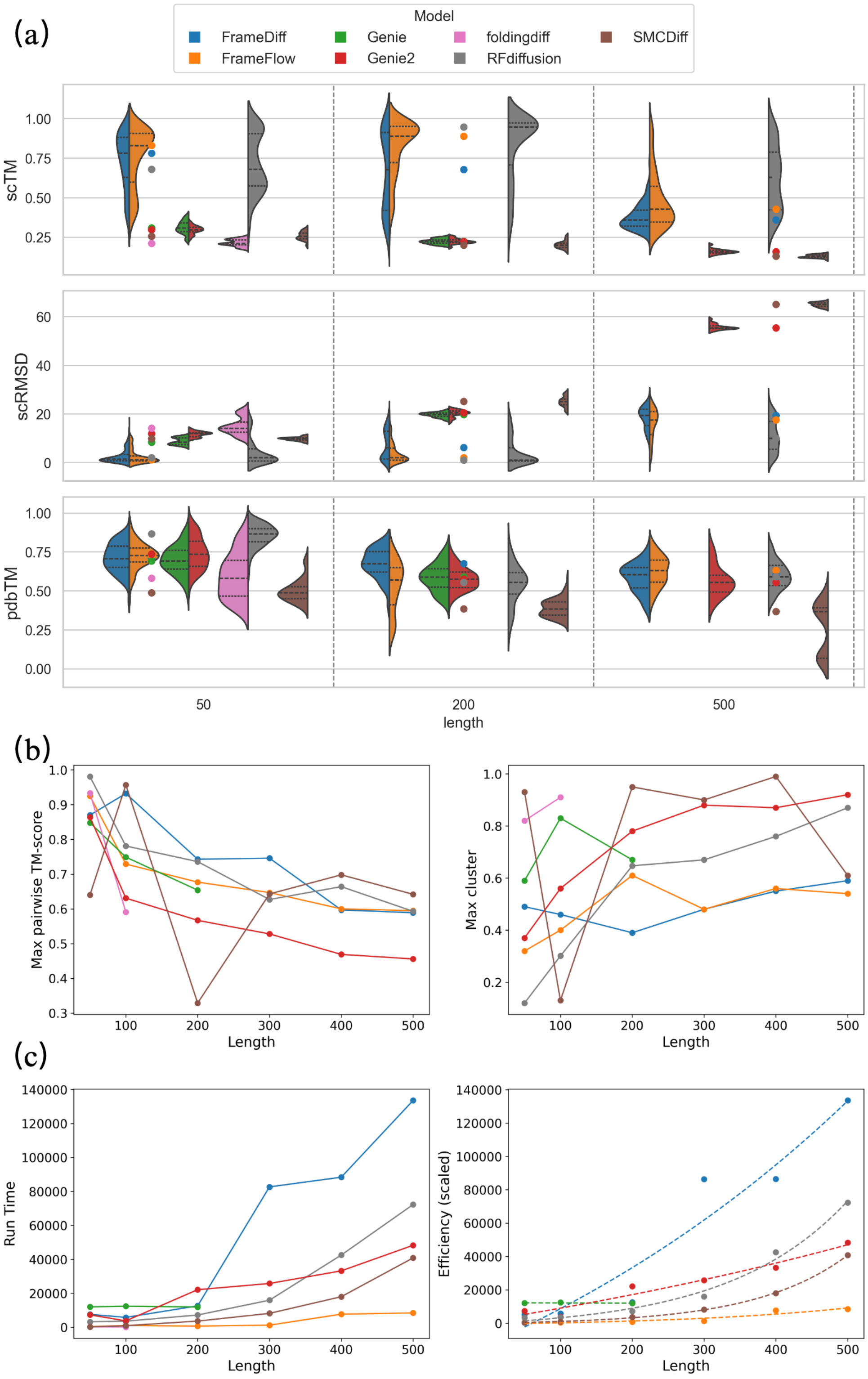

Figure 4. **Evaluation results on unconditional generation. (a) Results of designability (upper and middle parts) and novelty (lower part).** Three violin plots are presented, showing the distribution of TM-score (↑, the higher the better), scRMSD (↓), and pdbTM (↓). **(b) Results of diversity.** A line chart is used to illustrate the trend of pairwise TM (left), Max Cluster (right). **(c) Results of efficiency.** The runtime (left) is fitted to an exponential function (right) with respect to sequence length.