Unlocking Multimodal Protein Language Models at Inference Time

Yi Zhou[1], Qipeng Wang[1], Yunqing Liu[1], Jun Xia[2], Qing Li[1], Wenqi Fan[1,✉],
[1]The Hong Kong Polytechnic University,
[2]The Hong Kong University of Science and Technology (Guangzhou),
echo-yi.zhou@connect.polyu.hk; qipeng26.wang@connect.polyu.hk; yunqing617.liu@connect.polyu.hk;
junxia@hkust-gz.edu.cn; csqli@comp.polyu.edu.hk; wenqifan03@gmail.com

Abstract

Multimodal protein language models (pLMs) learn joint protein sequence-structure distributions, and their generation performance should also depend critically on inference-time sampling strategies. Yet prior work has focused more on model training than on how inference-time strategies behave. In this paper, we establish a three-stage investigation framework to empirically study the inference design space of multimodal pLMs across three representative pLMs and four fundamental tasks. We evaluate vanilla sampling, task-specific classifier-free guidance, and reward-guided beam search on multimodal pLMs, corresponding to controls over sampling distributions, per-step logits, and parallel trajectories. Throughout the complementary advancements centered on exploration-exploitation trade-off, we (1) reveal the suboptimality of default inference protocols and identify task-oriented sampling preferences; (2) observe substantial quantitative gains across tasks, consistently boosting the upper bound performance of multimodal pLMs without updating model parameters; (3) derive conclusions about base models that differ from prior consensus. Our code is available at: github.com/EchoChou990919/mplm_inference.

1 Introduction

Proteins are essential macromolecules in virtually all cellular processes. Their behavior follows the sequence-structure-function paradigm (Anfinsen, 1973) – the amino acid sequence determines the 3D structure, which in turn defines biological function. This biological centrality, combined with the inherent challenge of aligning heterogeneous data modalities within a unified semantic space, has driven extensive research on computational protein modeling (Fan et al., 2025; Notin et al., 2024; Liu et al., 2026, 2025b). Especially in the context of protein sequence-structure duality, key tasks like protein structure prediction (Jumper et al., 2021; Lin et al., 2023), inverse folding (Dauparas et al., 2022), motif scaffolding (Watson et al., 2023), and unconditional protein sequence-structure co-generation (Geffner et al., 2026) have attracted sustained research interest.

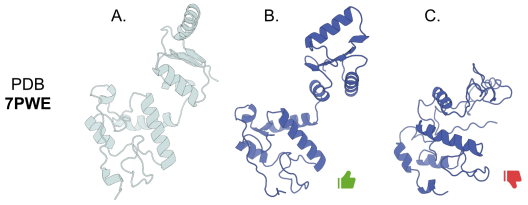

Figure 1: A structure prediction case. **A**. Ground-truth. **B**. One-step argmax prediction result. **C**. L-step temperature-based sampling result.

In recent years, multimodal generative protein language models (**pLMs**) have emerged as an "all-in-one" solution for **joint protein sequence-structure modeling** (Fan et al., 2025; LIU et al., 2026). Representative models, such as ESM3 (Hayes et al., 2025) and DPLM-2 series (Wang et al., 2025; Hsieh et al., 2025), treat protein sequence and structure as parallel token tracks: they learn the multimodal probability distribution of proteins during pre-training and address diverse tasks via iterative sampling from pLM logits at inference time. As with modern language and vision generative models (Karras et al., 2024; Snell et al., 2025), the performance of multimodal pLMs on generative tasks depends not only on the probability landscape learned but also on the sampling strategy employed, which shapes the inference-time sampling trajectory.

Figure 1 illustrates a protein structure prediction case comparing two typical inference-time sampling configurations. While single-step decoding at temperature $= 0.0$, ESM3 successfully predicts the structure of protein 7PWE in generally the same fold (TM-score > 0.5). However, when structure tokens are iteratively sampled at temperature $= 1.0$

✉Corresponding author: Wenqi Fan.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 35132–35151
October 24-29, 2026 ©2026 Association for Computational Linguistics

over L steps, ESM3's final prediction obviously collapses. This disparity intuitively underscores the significance of sampling strategies at inference time. On the same pLM landscape and starting from the same conditional point, inference over an exploitative short trajectory or an exploratory long trajectory yields different outputs.

Nevertheless, most research on multimodal pLMs has focused primarily on the effectiveness of training and fine-tuning, leaving inference-time sampling strategies largely underexplored. ESM3 (Hayes et al., 2025), DPLM-2 (Wang et al., 2025), and DPLM-2.1 (Hsieh et al., 2025) perform protein generation tasks by default using *model- and task-specific vanilla sampling configurations*, without the widely recognized *reward-free guidance* (Ho and Salimans, 2022; Karras et al., 2024; Nie et al., 2025) and *reward-guided search* (Yao et al., 2023; Wei et al., 2025) techniques. As a result, cross-model comparisons can be confounded by implementation-specific inference defaults. A few recent attempts apply advanced sampling to ESM3, e.g., a self-consistency reward for inverse folding (Liu et al., 2025a) or a foldability reward for unconditional generation (Li et al., 2025), but each is confined to one task, a single base model, or a specific strategy. We still lack a systematic understanding of *how multi-level sampling strategies behave across multimodal pLMs and tasks*, and *how the underlying exploration-exploitation trade-off should be navigated at inference time*.

In this paper, we address this gap through a systematic investigation that organizes **inference-time sampling** along three orthogonal control axes at different granularity levels: the sampling distribution, the per-step logits, and the parallel sampling trajectories, with an overview presented in Figure 2. **Section 3**: At the distribution level, we benchmark **vanilla sampling** among ESM3, DPLM-2, and DPLM-2.1 on four protein modeling tasks, sweeping the multimodal sampling order, total denoising step, unmasking, remasking, and temperature settings. **Section 4**: At the logit level, we instantiate task-specific variants of **classifier-free guidance (CFG)** that redirect each step's distribution toward the condition-aligned region of the sampling space. **Section 5**: At the trajectory level, we introduce a reward-guided **beam search** that selects across parallel sampling trajectories using global quality signals. Our main contributions are summarized as follows, while the main insights are summarized at the end of each section:

- We introduce a systematic investigation framework for inference-time sampling in multimodal pLMs, orthogonally composing vanilla sampling, reward-free guidance, and reward-guided search strategies, and comprehensively covering four generative protein modeling tasks.
- The three-level inference strategies we investigate consistently advance the best empirical performance for multimodal pLMs across four tasks, especially boosting breakthroughs in ESM3's unconditional protein sequence-structure cogeneration and motif-scaffolding.
- We demonstrate that standardized inference-time strategies can substantially change conclusions about multimodal pLMs: models viewed as weak under default sampling can become competitive once inference strategies are controlled, revealing inference protocol as a hidden confounder in protein modeling evaluation.

2 Preliminaries

Multimodal Protein Language Modeling: Formulation and Benchmarks. A protein is fundamentally characterized by its sequence and structure. We study multimodal pLMs, represented by ESM3 (Hayes et al., 2025) and DPLM-2 series (Wang et al., 2025; Hsieh et al., 2025), which model proteins via parallel token tracks. For a protein of length L, its sequence is represented as $\boldsymbol{s} = (s_1, \ldots, s_L)$ over the standard 20-amino-acid vocabulary, and its structure as a discrete latent sequence $\boldsymbol{z} = (z_1, \ldots, z_L)$ produced by the model's structure tokenizer. These pLMs learn the joint distribution $p_\theta(\boldsymbol{s}, \boldsymbol{z})$ with neural network θ, from which four multimodal protein generative tasks can be conducted: unconditional sequence-structure co-generation via $p_\theta(\boldsymbol{s}, \boldsymbol{z}|\varnothing)$, motif scaffolding via $p_\theta(\boldsymbol{s}, \boldsymbol{z}|\boldsymbol{s}_{\text{motif}}, \boldsymbol{z}_{\text{motif}})$, structure prediction via $p_\theta(\boldsymbol{z}|\boldsymbol{s})$, and inverse folding via $p_\theta(\boldsymbol{s}|\boldsymbol{z})$.

Our study focuses on investigating inference protocols for multimodal pLMs. Regarding the evaluation datasets and metrics, we follow established benchmarks. For unconditional co-generation, we generate protein samples ranging from 100 to 500 residues and report designability and diversity metrics (Geffner et al., 2026). Motif scaffolding is evaluated on 24 benchmark problems (Zheng et al., 2025), with the number of solved problems, overall success rate, and diversity of successful designs reported (Wang et al., 2025; Yim et al., 2024). Protein structure prediction and inverse folding

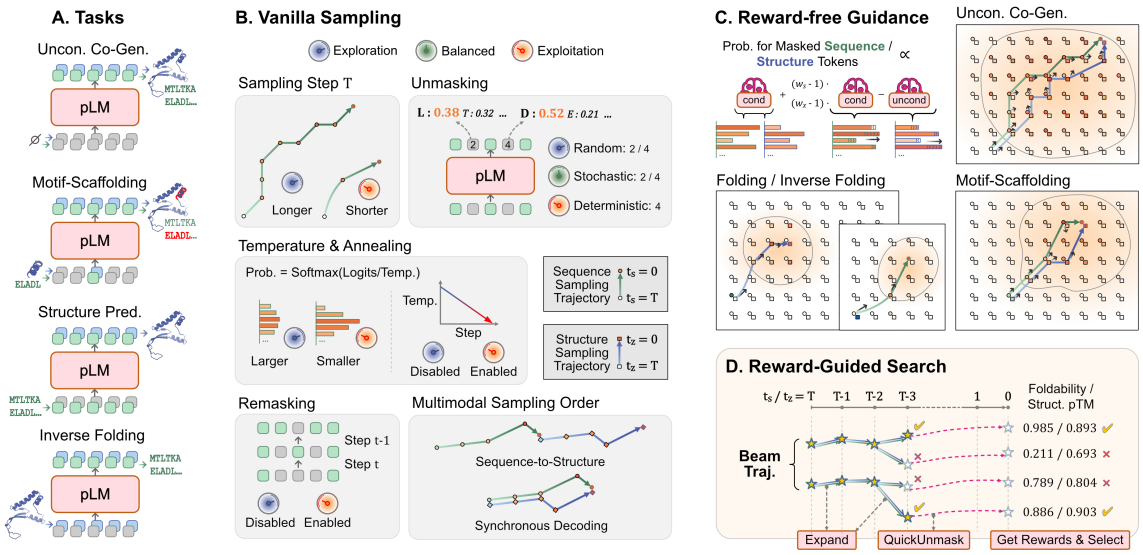

Figure 2: **Overview of our empirical investigation framework. A.** Multimodal protein language models (pLMs) perform four foundational protein modeling tasks via iterative diffusion sampling. **B.** At the distribution level, we benchmark **vanilla sampling** across three pLMs and four tasks, therefore identifying the task-oriented exploration-exploitation trade-off at inference time. **C.** At the logit level, we propose **task-motivated classifier-free guidance** that steers each step's logits and therefore directs the sampling trajectory. **D.** At the trajectory level, **reward-guided beam search** navigates parallel trajectories using model-internal quality signals, highlighting the effectiveness of "explore-then-select" inference across tasks.

are evaluated on the CAMEO2022 and PDB date datasets (Campbell et al., 2024), with RMSD & TM-score, and scTM & AAR reported (Wang et al., 2025; Hsieh et al., 2025). Additionally, we record the FLOPs and wall-clock runtime to analyze the computational efficiency of inference protocols. Details of the protein structure tokenization, base models, and evaluation pipelines are deferred to Appendix A.1.

Protein Generation with Diffusion Language Models. Typical multimodal pLMs (Hayes et al., 2025; Wang et al., 2025; Hsieh et al., 2025) are built on the discrete diffusion framework. At inference, protein generation proceeds along a reverse denoising trajectory of T **steps**, iteratively refining an initial masked state $(\boldsymbol{s}^{(T)}, \boldsymbol{z}^{(T)})$ into a clean state $(\boldsymbol{s}^{(0)}, \boldsymbol{z}^{(0)})$. With two token tracks, multimodal protein generation admits different **sampling orders** along the two diffusion timelines t_s and t_z: a *sequence-to-structure* schedule completes t_s before advancing t_z, while a *synchronous* schedule couples them under a shared timeline ($t_s = t_z = t$).

Unlike left-to-right autoregressive decoding, each diffusion step involves two sets of operations on the masked positions. First, an explicit **unmasking** policy selects an unmasking subset $\mathcal{U}_t \subseteq \mathcal{M}_t$ from the currently masked indices $\mathcal{M}_t$, based on a per-position confidence $c_i^{(t)}$ (e.g., log-probability or entropy), in one of three modes: *deterministic* (best $c_i^{(t)}$ first), *stochastic* (Gumbel-perturbed confidence), or *random*. An optional **remasking** action re-masks low-confidence tokens for later refinement. Second, for each $i \in \mathcal{U}_t$, the token is sampled from a **temperature**-scaled categorical distribution $p_i^{(t)}(\cdot) = \mathrm{softmax}(\mathbf{l}_i^{(t)}/\tau_t)$ over raw logits $\mathbf{l}_i^{(t)}$. Larger τ_t increases randomness, while a smaller one makes decoding greedier. The temperature τ_t can also be **annealed** across steps to reduce entropy and stabilize the final state.

Reward-free Guidance and Reward-Guided Search. Reward-free guidance, particularly classifier-free guidance (CFG) (Ho and Salimans, 2022; Karras et al., 2024), is a standard practice in diffusion vision models that steers sampling toward context-aligned high-likelihood regions. Prior work has proven that CFG can be applied to diffusion language models with a straightforward adaptation (Nie et al., 2025; Schiff et al., 2025; Chen et al., 2025). At each step t, the guided logits are computed as $\mathbf{l}_{\text{guided}}^{(t)} = \mathbf{l}_{\text{uncond}}^{(t)} + w \cdot (\mathbf{l}_{\text{cond}}^{(t)} - \mathbf{l}_{\text{uncond}}^{(t)})$, where $\mathbf{l}_{\text{cond}}^{(t)}$ and $\mathbf{l}_{\text{uncond}}^{(t)}$ are the conditional and unconditional logits and w is the guidance scale, with $w > 1$ amplifying the conditional signal and sharpening adherence to the specified constraints.

Beyond single-trajectory guidance, search algorithms (Snell et al., 2025; Wei et al., 2025) use a global reward with two complementary views: *exploring multiple trajectories in parallel*, as in Best-of-N, and *evaluating the ongoing trajectory in advance*, as in soft value-based decoding (SVDD) (Li et al., 2025). Beam search can be generalized to

combine both views (Didi et al., 2026), which we treat in a unified manner in Section 5.

3 Vanilla Sampling

To begin our investigation of the inference-time behavior of multimodal pLMs, we revisit the default sampling of ESM3, DPLM-2, and DPLM-2.1 across four foundational tasks, with details provided in Appendix A.2. Two issues motivate a closer look. First, the inference implementations across models are not aligned: ESM3 lacks native support for stochastic unmasking, remasking, and synchronous sequence-structure sampling, which are integral to DPLM-2 and DPLM-2.1. Second, across tasks, the four sets of defaults rely on coarse heuristics, whose rationality and optimality have not been empirically verified. Together, these discrepancies hinder fair cross-model comparison and motivate a unified sampling framework.

To address these limitations, **we standardize inference implementations across multimodal pLMs and benchmark vanilla sampling configurations**. Specifically, we conduct a comprehensive grid search across six key dimensions: diffusion steps T (spanning 1, 8, $L/8$, $L/4$, $L/2$, L, and optionally 100 and 500), unmasking strategies (deterministic, stochastic, or random), remasking (enabled or disabled), sampling temperature (varying in $[0, 1]$), temperature annealing (enabled or disabled), and multimodal sampling orders (synchronous, or sequence-to-structure). The resulting optimal configurations differ considerably from the defaults across all three models and four tasks.

Figure 3 and Appendix A.2.2-A.2.3 present the optimal configurations and ablations for vanilla sampling settings. Taking ESM3 as representative, on the unconditional co-generation benchmark, enabling synchronous multimodal sampling, stochastic unmasking, and remasking unlocks ESM3 from subpar to considerably competitive. Under the default sequence-to-structure sampling, ESM3 produces only 52.8 designable protein samples and 34.4 unique designable clusters. With the optimized configuration, it reaches 174.8 designable samples and 64.4 clusters, on par with DPLM-2 (149.0 designable samples and 76.0 clusters). On the motif-scaffolding benchmark, ESM3 solves obviously more problems, increasing from an average of 19.4 to 21.6 out of 24. On the CAMEO2022 dataset, optimizing the sampling configuration yields consistent gains: the average TM-score for protein structure prediction rises from 0.860 to 0.868, and the average scTM for inverse folding improves from 0.899 to 0.905.

The optimal configurations and extent of gains differ sharply across tasks, motivating a unified, task-aware view of inference behavior through **the exploration-exploitation trade-off**. The trade-off should be governed by the *length of the denoising trajectory* and the *randomness introduced at each step* and *accumulated over steps*, along which the sampling distribution is iteratively shaped. Our six grid-search dimensions provide concrete controls over these factors: total steps T determine trajectory length; unmasking strategy and sampling temperature control per-step randomness; remasking and temperature annealing affect randomness accumulation; and the sequence-structure sampling order acts as a cross-track control in multimodal protein generation. The corresponding exploration or exploitation tendency for each setting is summarized in Figure 2.B. Reading the optimized configurations through this lens, the four protein modeling tasks fall into distinct regions of the trajectory-by-randomness plane, consistent with the topology of their task-specific solution spaces.

- † **Unconditional co-generation targets the full protein space and lies near the exploration corner**, requiring the longest trajectories with stochastic unmasking, remasking, and high temperatures, while relying on synchronous decoding to enforce protein designability.
- † **Motif scaffolding occupies the balanced regime**, where the solution space must simultaneously satisfy motif constraints, scaffold self-consistency, and diversity, calling for long trajectories, moderate temperature, and synchronous sequence-structure decoding.
- † **Protein structure prediction and inverse folding sit near the exploitation corner**. The tight biophysical coupling between sequence and structure yields a narrow but non-degenerate solution space, favors short trajectories, deterministic or stochastic unmasking, and low temperatures.

4 Reward-Free Guidance

Classifier-free guidance (CFG) (Ho and Salimans, 2022) has been widely acknowledged in general generative modeling, yet remains underexplored in multimodal pLMs. In our investigation framework,

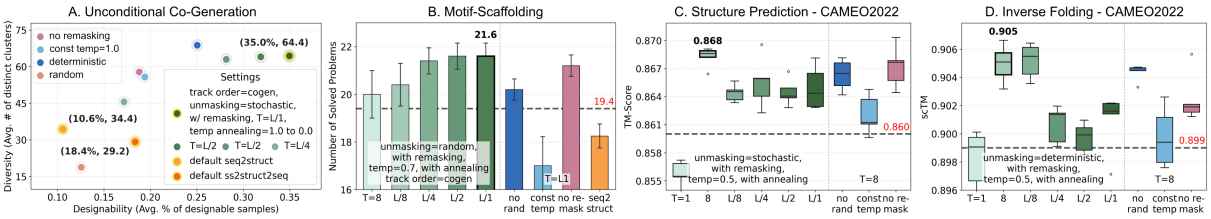

Figure 3: **Vanilla sampling benchmark**. Panels **A-D** show ESM3 performance on each task under the optimal vanilla sampling configuration, with ablations over total steps, per-step randomness (unmasking $\times$ temperature), temperature annealing, remasking, and multimodal sampling order. The optimal configuration is highlighted, and the default performance is marked for reference.

CFG adds an orthogonal control axis to the vanilla sampling strategies. While vanilla sampling shapes the foundational sampling distribution, CFG instead steers the per-step logits, shifting probability mass away from the unconditional prior and toward the condition-aligned region of the solution space. In this section, **we design CFG strategies for four multimodal protein modeling tasks, implement them across multimodal pLMs, and examine how this logit-level control behaves across task regimes and model capabilities**. To accommodate the dual-track nature of multimodal pLMs, we extend standard CFG to a track-wise form. Let $m \in \{s, z\}$ index the sequence and structure tracks, the guided logits on each track are then

$$\mathbf{l}^{m,(t)}_{\text{guided}} = \mathbf{l}^{m,(t)}_{\text{uncond}} + w_m \cdot \left(\mathbf{l}^{m,(t)}_{\text{cond}} - \mathbf{l}^{m,(t)}_{\text{uncond}}\right), \quad (1)$$

where w_m is the track-wise guidance scale. Each subsection below instantiates this formulation by specifying (i) which tracks are denoised and (ii) the conditional and unconditional forward inputs that yield $\mathbf{l}^{m,(t)}_{\text{cond}}$ and $\mathbf{l}^{m,(t)}_{\text{uncond}}$. The CFG methodology and schematic task-specific guided trajectories are also sketched in Figure 2.C. All experiments build on the optimal vanilla sampling strategies, with selected CFG scales reported in Appendix A.3.1.

Unconditional Protein Sequence-Structure Co-Generation. In unconditional co-generation, multimodal pLMs receive no explicit conditioning but must maintain protein sequence-structure consistency throughout the denoising. Our earlier observation demonstrates that track-by-track sampling underperforms synchronous sequence-structure decoding, in which each sampling step can be viewed as a cross-modal update, with the partially decoded tokens on one track implicitly conditioning the other. This perspective motivates using CFG to amplify the cross-modal conditioning signal at each step. We thus formulate CFG as:

$$\mathbf{l}^{s,(t)}_{\text{cond}}, \mathbf{l}^{z,(t)}_{\text{cond}} = f_\theta(\boldsymbol{s}^{(t)}, \boldsymbol{z}^{(t)}),$$
$$\mathbf{l}^{s,(t)}_{\text{uncond}} = f_\theta(\boldsymbol{s}^{(t)}, \varnothing), \quad \mathbf{l}^{z,(t)}_{\text{uncond}} = f_\theta(\varnothing, \boldsymbol{z}^{(t)}),$$

where each unconditional logit is computed by masking the tokens on the opposite track. Guided logits on each track are then computed using Equation 1 with track-specific scales w_s and w_z.

As shown in Table 1, **cross-modal CFG improves the generation designability and the diversity across all three base models**. For DPLM-2 and DPLM-2.1, the size of the designable subset and the number of distinct designable clusters increase moderately over their default sampling. On ESM3, the gains are much more pronounced. Across all generated samples, ESM3 with CFG attains better sequence foldability (average pLDDT of 88.748) and stronger sequence-structure self-consistency (average scTM of 0.931). The number of designable protein samples reaches 325.0, nearly doubling that of the optimal vanilla sampling (174.8), while the number of distinct designable clusters rises to 116.6, about 1.8 times that of the optimal vanilla sampling (64.4). In prior protein co-design studies, ESM3 has commonly been regarded as an underperforming baseline. Our results show that **once equipped with optimal vanilla sampling and a properly designed CFG, ESM3 delivers a strong performance**, surpassing both DPLM-2 and DPLM-2.1 by a clear margin.

Beyond the quantitative gains, the qualitative behavior of ESM3 also changes. Under default sampling, ESM3 often produces implausible proteins with repetitive residues and a biased secondary-structure composition. Then, **cross-modal CFG can make the generations more "protein-like".** As shown in Figure 4, applying our cross-modal CFG reduces residue repetition and brings the α-helix/β-sheet/coil proportions closer to those of natural proteins. In addition, Appendix A.3.2 presents a PDB alignment analysis that further demonstrates that cross-modal CFG makes the generations structurally more similar to PDB proteins. We interpret this qualitative effect as an implicit regularization of joint multimodal sampling. By up-weighting cross-track conditional logits, CFG discourages the

Table 1: Evaluation of Unconditional Sequence-Structure Co-Generation. Note: Improves over the last investigation stage.

Model	Cost*	Designable Subset		All Samples			
		#Design ↑	#Clusters ↑	pLDDT ↑	scTM ↑	#Clusters ↑	α / β %
DPLM-2 $_{\text{Default}}$	2.18, 1.17	149.0 (126, 160)	76.0 (62, 84)	81.920 ± 8.643	0.906 ± 0.105	262.0 (248, 273)	38.4 / 16.9
DPLM-2 $_{\text{CFG}}$	6.54, 3.22	162.4 (157, 168)	78.4 (75, 84)	82.242 ± 8.247	0.913 ± 0.091	247.4 (241, 254)	36.9 / **17.4**
DPLM-2.1 $_{\text{Default}}$	2.15, 1.02	143.4 (129, 153)	86.6 (82, 97)	85.116 ± 7.681	0.904 ± 0.106	304.2 (295, 312)	43.6 / 13.6
DPLM-2.1 $_{\text{CFG}}$	6.44, 3.00	162.2 (145, 171)	96.4 (90, 104)	84.995 ± 7.671	0.911 ± 0.099	284.8 (272, 293)	42.5 / 14.2
ESM3 $_{\text{Default seq2struct}}$	1.25, 1.68	52.8 (31, 70)	34.4 (24, 49)	61.387 ± 17.563	0.660 ± 0.242	**381.4 (371, 396)**	48.3 / 4.3
ESM3 $_{\text{Default ss2struct2seq}}$	3.75, 4.98	91.8 (90, 95)	29.2 (22, 45)	76.079 ± 13.530	0.762 ± 0.221	240.0 (229, 260)	71.3 / 3.5
ESM3 $_{\text{Vanilla}}$	1.25, 1.87	174.8 (167, 179)	64.4 (56, 69)	68.724 ± 26.021	0.683 ± 0.294	135.0 (124, 148)	43.5 / 5.3
ESM3 $_{\text{CFG}}$	3.75, 2.74	325.0 (315, 333)	116.6 (106, 123)	88.748 ± 12.007	0.931 ± 0.145	194.2 (183, 202)	53.8 / 10.4
ESM3 $_{\text{Beam - Select=Best}}$	-	**411.6 (406, 421)**	126.6 (122, 131)	**91.577 ± 5.208**	**0.976 ± 0.044**	148.0 (142, 158)	45.8 / 12.4
ESM3 $_{\text{Beam - Select=Random}}$	15.4, 11.88	350.4 (336, 360)	**139.0 (132, 145)**	90.704 ± 5.525	0.964 ± 0.056	177.6 (170, 186)	47.9 / 11.6
ESM3 $_{\text{Official SVDD}}$	-	189.4 (175, 201)	88.4 (79, 94)	76.478 ± 15.430	0.881 ± 0.160	253.4 (241, 266)	51.7 / 10.4
La-Proteina	-	383.0 (370, 397)	134.7 (125, 141)	83.770 ± 10.130	0.953 ± 0.119	206.0 (202, 208)	72.1 / 5.7

* denotes the computational cost at inference time, concretely the "FLOPs ($\times 10^{17}$), Runtime (Hours)".

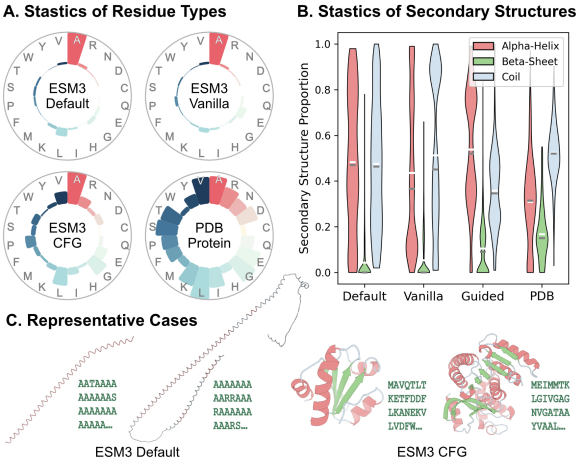

Figure 4: **Qualitative Analysis. A-B.** Optimal vanilla sampling and CFG alleviate the abnormal properties of ESM3-generated proteins. **C.** Visualizations of representative cases.

tendency for one track to move toward plausible states in isolation while having low joint probability with the other track. Prior work sought to mitigate the repetition-induced generation collapse in pLMs by injecting external steering vectors (Zhang et al., 2026), yet our results show that internal cross-modal steering can already be highly effective.

Motif-Scaffolding. In motif-scaffolding, multi-modal pLMs sample both sequence and structure tracks to generate coherent protein scaffolds for specific motifs. Success requires both preserving local motifs and maintaining global protein designability. The former emphasizes the explicit conditioning of the motifs. The latter, along with our observation of the superiority of synchronous decoding over sequence-to-structure sampling, highlights the implicit cross-modal conditioning. We therefore use a track-wise CFG that emphasizes both motif and cross-modal alignments. The conditional logits on both tracks are obtained from $\mathbf{l}^{s,(t)}_{\text{cond}}, \mathbf{l}^{z,(t)}_{\text{cond}} = f_\theta(\boldsymbol{s}^{(t)}, \boldsymbol{z}^{(t)})$. The unconditional logits are computed by removing motif and cross-modal information from each track:

$$\mathbf{l}^{s,(t)}_{\text{uncond}} = f_\theta\big(\boldsymbol{s}^{(t)} \setminus \boldsymbol{s}_{\text{motif}}, \varnothing\big),$$
$$\mathbf{l}^{z,(t)}_{\text{uncond}} = f_\theta\big(\varnothing, \boldsymbol{z}^{(t)} \setminus \boldsymbol{z}_{\text{motif}}\big).$$

The guided logits on each track then follow Equation 1 with track-specific scales w_s and w_z.

Table 2: Evaluation of Motif-Scaffolding.

Model	Cost*	#Solved ↑	Succ. Rate ↑	#Clusters ↑
DPLM-2 $_{\text{Default}}$	2.88, 1.37	15.6 (15, 16)	15.3 ± 0.3 %	72.6 (63, 83)
DPLM-2 $_{\text{Vanilla}}$	0.47, 0.42	16.8 (16, 17)	20.6 ± 0.5 %	72.8 (68, 80)
DPLM-2 $_{\text{CFG}}$	1.41, 0.75	15.8 (15, 17)	20.4 ± 0.2%	71.6 (63, 81)
DPLM-2.1 $_{\text{Default}}$	2.83, 1.15	14.2 (14, 15)	18.7 ± 0.6 %	51.0 (45, 55)
DPLM-2.1 $_{\text{Vanilla}}$	2.83, 1.15	15.4 (15, 16)	19.4 ± 0.6 %	66.8 (61, 72)
DPLM-2.1 $_{\text{CFG}}$	8.49, 3.33	15.0 (14, 16)	19.3 ± 0.5 %	61.2 (56, 68)
ESM3 $_{\text{Default}}$	0.18, 1.11	19.4 (19, 20)	19.6 ± 0.8 %	94.2 (85, 98)
ESM3 $_{\text{Vanilla}}$	0.28, 1.66	21.6 (21, 22)	30.9 ± 1.1 %	110.0 (100, 124)
ESM3 $_{\text{CFG}}$	0.84, 1.67	22.8 (22, 23)	37.5 ± 1.0 %	179.4 (165, 198)
ESM3 $_{\text{Beam}}$	3.65, 15.64	**23.0 (23, 23)**	**49.9 ± 0.7 %**	**261.0 (250, 278)**
La-Proteina $_{\text{scaff}}$	-	22.0 (22, 22)	42.7 ± 0.7 %	256.6 (247, 272)

* Inference cost in "FLOPs ($\times 10^{17}$), Runtime (Hours)".

Table 2 summarizes the results. **While CFG brings no further gains over optimal vanilla sampling for DPLM-2 and DPLM-2.1, it substantially improves ESM3**. In 4 out of 5 independent runs, ESM3 solves 23 of the 24 benchmark cases, outperforming DPLM-2/-2.1, and ESM3 itself without guidance. The average success rate rises to 37.5%, nearly doubling that of the default sampling and about 1.2 times that of the optimal vanilla sampling. The diversity of successful scaffolds is improved accordingly: ESM3 with CFG produces 179.4 unique solution clusters on average, nearly twice that of the default sampling (94.2) and about 1.6 times that of the optimal vanilla sampling (110.0). We interpret this contrast as a difference in **how reliably each base model encodes the motif and cross-modal conditions**. ESM3 appears to encode these conditions strongly enough for effective logit-level steering, whereas DPLM-2 and DPLM-2.1 do not. We further provide an ablation on the dual-condition CFG design in Appendix A.3.3.

Table 3: Evaluation of Protein Structure Prediction (Left) and Inverse Folding (Right).

Model	CAMEO			PDB Date Split		Model	CAMEO			PDB Date Split	
	Cost*	RMSD ↓	TM-score ↑	RMSD ↓	TM-score ↑		Cost*	scTM ↑	AAR	scTM ↑	AAR
DPLM-2 $_{\text{Default}}$	18.5, 6.2	7.483±6.126	0.786±0.170	5.253±5.143	0.836±0.144	DPLM-2 $_{\text{Default}}$	18.5, 6.2	0.870±0.158	48.1	0.912±0.111	55.6
DPLM-2 $_{\text{Vanilla}}$	5.6, 2.8	7.513±6.178	0.790±0.167	5.249±5.103	0.837±0.141	DPLM-2 $_{\text{Vanilla}}$	0.2, 0.1	0.875±0.153	48.2	0.917±0.111	55.6
DPLM-2 $_{\text{CFG}}$	11.2, 4.1	7.414±6.112	0.786±0.169	5.178±5.029	0.835±0.142	DPLM-2 $_{\text{CFG}}$	0.4, 0.1	0.877±0.151	47.1	0.922±0.101	54.9
DPLM-2.1 $_{\text{Default}}$	18.2, 5.5	6.297±6.181	0.822±0.170	-	-	DPLM-2.1 $_{\text{Default}}$	18.2, 5.5	0.875±0.144	53.3	-	-
DPLM-2.1 $_{\text{Vanilla}}$	1.5, 0.5	6.259±6.051	0.824±0.167	-	-	DPLM-2.1 $_{\text{Vanilla}}$	1.5, 0.5	0.879±0.142	53.4	-	-
DPLM-2.1 $_{\text{CFG}}$	2.9, 0.9	6.104±5.809	0.824±0.167	-	-	DPLM-2.1 $_{\text{CFG}}$	2.9, 0.9	0.880±0.148	52.1	-	-
ESM3 $_{\text{Default}}$	0.1, 0.3	5.334±6.280	0.860±0.168	4.053±4.867	0.882±0.152	ESM3 $_{\text{Default}}$	0.8, 1.4	0.899±0.143	46.4	0.944±0.083	49.6
ESM3 $_{\text{Vanilla}}$	0.8, 1.2	5.563±7.108	0.868±0.161	4.872±7.178	0.885±0.150	ESM3 $_{\text{Vanilla}}$	0.8, 1.4	0.905±0.134	48.4	0.947±0.085	52.0
ESM3 $_{\text{CFG}}$	1.6, 1.3	5.144±6.129	0.870±0.157	3.969±4.823	0.889±0.142	ESM3 $_{\text{CFG}}$	1.6, 1.5	0.910±0.135	45.7	0.954±0.075	49.1
ESM3 $_{\text{Beam}}$	35.6, 36.2	5.221±6.565	0.872±0.157	3.744±4.542	0.893±0.140	ESM3 $_{\text{Beam}}$	47.4, 50.0	**0.914±0.129**	46.0	**0.956±0.069**	49.3
ESMFold	-	**3.961±4.795**	**0.898±0.147**	**2.661±3.916**	**0.931±0.112**	ProteinMPNN	-	0.902±0.145	42.4	0.953±0.075	46.8

* Inference cost in "FLOPs ($\times 10^{15}$), Runtime (Minutes)".

Protein Structure Prediction. For structure prediction, multimodal pLMs condition on the input sequence $\boldsymbol{s}$ and perform iterative discrete diffusion sampling along the structure track $\boldsymbol{z}$. At each denoising step t, the conditional structure logits are obtained through a forward pass $\mathbf{l}^{z,(t)}_{\text{cond}} = f_\theta(\boldsymbol{s}, \boldsymbol{z}^{(t)})$. The corresponding unconditional logits are obtained by fully masking the sequence track and executing an additional forward pass, yielding $\mathbf{l}^{z,(t)}_{\text{uncond}} = f_\theta(\varnothing, \boldsymbol{z}^{(t)})$. Guided sampling is then performed by extrapolating the two via Equation 1. Table 3 summarizes the evaluation results.

For DPLM-2 and DPLM-2.1, CFG yields modest RMSD improvements and no consistent TM-score gains. In contrast, **ESM3 benefits clearly from CFG across both datasets and metrics**, achieving an average RMSD of 5.144 and TM-score of 0.870 on CAMEO2022, and an RMSD of 3.969 with TM-score of 0.889 on PDB Date. We attribute this contrast to **the narrow solution space in structure prediction and the differing sequence-conditional capacity across base models**. As a given sequence admits only a small set of plausible folds, CFG should be useful when a pLM has not yet fully exploited its sequence condition. DPLM-2 and DPLM-2.1 already operate near their conditional capacity under optimal vanilla sampling, whereas ESM3, with the strongest capability of the three, leaves more room above its conditional optimum for logit-level steering. Additionally, a divergence between RMSD and TM-score performance in ESM3 is analyzed in Appendix A.3.4.

Inverse Folding. For inverse folding, multimodal pLMs iteratively sample along the sequence track $\boldsymbol{s}$, conditioned on the target structure $\boldsymbol{z}$. We formulate CFG for this task symmetrically to structure prediction. At each denoising step t, the conditional sequence logits are $\mathbf{l}^{s,(t)}_{\text{cond}} = f_\theta(\boldsymbol{s}^{(t)}, \boldsymbol{z})$. The corresponding unconditional logits are obtained by dropping the structural condition and executing an additional forward pass, yielding $\mathbf{l}^{s,(t)}_{\text{uncond}} = f_\theta(\boldsymbol{s}^{(t)}, \varnothing)$.

Table 3 shows that **CFG consistently improves structural consistency and reduces sequence similarity to the native reference across base models**. For ESM3, the average scTM reaches 0.910 on CAMEO2022 and 0.954 on PDB Date, while the AAR declines slightly. We interpret this decline in light of the one-to-many nature of inverse folding and the exploitation effect of CFG. It improves structural consistency by shifting pLM probability mass toward protein sequences that better satisfy the target structure, without requiring them to recover the native reference.

Inference Efficiency. Alongside benchmark performance, Tables 1-3 also report the computational costs across inference protocols, revealing the inefficiency of default sampling and the superior performance-efficiency balance of our CFG strategies. See Appendix A.5 for detailed discussions.

† **CFG provides a simple and general control for multimodal protein modeling**. By casting diverse tasks into conditional generation and dropping the appropriate conditions in the unconditional branch, it consistently improves the protein generation performance.

† **CFG helps more when the task condition leaves a larger solution space, giving more room to steer the sampling**. It is evidenced by the larger gains on unconditional cogeneration and motif-scaffolding tasks, which have a considerable exploratory nature.

† **The benefit of CFG is model-dependent and scales with how well the base model has internalized the corresponding conditional signal.** Stronger multimodal pLMs expose more usable condition-aligned logit changes for steering, tend to gain more from CFG.

5 Reward-Guided Search

Our study of vanilla sampling and reward-free guidance shows that the performance of multimodal pLMs can be improved by optimizing the sampling trajectory. However, they still generate only a single trajectory per run and rely on per-step token-level logits, which assess local probabilities rather than the global quality of the final output, leaving potentially superior trajectories unexplored. To address this limitation, **we lift control from individual steps to entire trajectories: multiple trajectories are maintained in parallel and selected using global rewards.** Notably, this section focuses exclusively on ESM3. Our earlier results show that it is the only base model with consistent room for logit-level steering, making it the natural base for studying trajectory-level search. Implementation details are presented in Appendix A.4.

Specifically, we adapt beam search for the inference of multimodal pLMs, as illustrated in Figure 2.D and Algorithm 1, and as explained in detail in Appendix A.4.2. Built on top of the underlying sampler, the search layer is decoupled from vanilla sampling and CFG choices.

Let $\mathcal{T}^{(t)}$ denote the retained sampling trajectories at step t, where each trajectory is a partially denoised protein $(\boldsymbol{s}^{(t)}, \boldsymbol{z}^{(t)})$. Starting from $\mathcal{T}^{(T)}$ with beam width N, branching factor B, and scoring interval K, we perform expand-score-select operations periodically:

$$\mathcal{C}^{(t)} \leftarrow \textsc{Expand}(\mathcal{T}^{(t)}, B), \quad \text{if } t \bmod K = 0,$$
$$\hat{r}(\boldsymbol{c}) \leftarrow r(\textsc{QuickUnmask}(\boldsymbol{c})), \quad \boldsymbol{c} \in \mathcal{C}^{(t)},$$
$$\mathcal{T}^{(t-1)} \leftarrow \textsc{Select}(\mathcal{C}^{(t)}, \hat{r}, N).$$

Every K steps (when $t \bmod K = 0$), the search maintains N parallel trajectories and expands each trajectory into B candidates via independent one-step samplings. While these candidates are still partially masked, we use the estimated reward $\hat{r}$: each candidate is denoised by one-pass $\textsc{QuickUnmask}(\cdot)$ and then scored by a reward $r(\cdot)$. The search then prunes the beam set back to width N under $\hat{r}$ via a selection rule $\textsc{Select}(\cdot)$.

At non-expansion steps (when $t \bmod K \neq 0$), all retained trajectories are simply advancing. After the final step, a single output is selected from the beam set $\mathcal{T}^{(0)}$ using $\textsc{Select}(\cdot)$ again.

Within this framework, task-specific design reduces to the reward function $r(\cdot)$ and the selection rule $\textsc{Select}(\cdot)$. To preserve the "all-in-one" nature of multimodal pLMs and keep comparisons with vanilla sampling and CFG fair, we use model-internal signals as rewards, i.e., pTM scores produced during protein structure de-tokenization. The structure track is scored directly by structural pTM, whereas the sequence track is first greedily folded into structure tokens and then scored, named foldability pTM. Then, the selection rule should match each task's exploration-exploitation profile: for protein structure prediction, inverse folding, and motif-scaffolding, $\textsc{Select}(\cdot)$ returns the top candidates ranked by structural pTM, foldability pTM, or their sum, respectively. For unconditional co-generation, we instead randomly select candidates that satisfy structural and foldability constraints (pTM > 0.8 on both tracks), thereby preserving diversity while enforcing a quality threshold.

Tables 1-3 also report the performance of ESM3 with reward-guided search, alongside specialist baselines. **Across all four tasks, beam search consistently improves over the CFG stage and enables a single foundation protein model to match or surpass task-specific SOTA systems without task-specific training or external rewards.** For unconditional co-generation, ESM3 attains the strongest designability across all baselines. The threshold-based random selection variant raises the number of unique designable clusters from 116.6 to 139.0, surpassing La-Proteina (Geffner et al., 2026). We also include the official SVDD implementation in ESM3, which works atop the default sequence-to-structure sampling. Its limited performance further suggests that our improvements from vanilla sampling and CFG are complementary to reward-guided search. On motif-scaffolding, ESM3 solves 23 of 24 problems in every run, raising the success rate from 37.5% to 53.3% and the number of unique clusters from 179.4 to 245.2, exceeding La-Proteina's motif-scaffolding variant. For protein structure prediction, ESM3 further reduces the average RMSD and lifts the TM-score on PDB Date, narrowing the gap to ESMFold (Lin et al., 2023). For inverse folding, ESM3 achieves the best scTM on both CAMEO2022 and PDB Date (0.914 and 0.956), surpassing ProteinMPNN (Dauparas et al., 2022).

Notably, beam search delivers its full benefit when the underlying vanilla sampling is tuned to be slightly more exploratory than its single-trajectory optimum, as leaving room for reward-driven selection to discriminate among candidate trajectories. Details are deferred to Appendix A.4.4. Besides, as transparently presented in Table 1-3, beam

search incurs considerable computational overhead at inference to achieve improved benchmark performance. Computational efficiency issues are discussed in detail in Appendix A.5.

† **Multimodal pLMs can support a simple and effective internal loop at inference time.** Searching across parallel trajectories with model-internal rewards further improves protein modeling performance without additional training or external judges.

† **Our three-stage evaluation pushes multimodal pLMs to the level of task-specific SOTA systems across multiple protein modeling tasks**. This may reshape the community's view of multimodal pLMs' position, which has so far been formed by their performance under default sampling.

6 Conclusion

In this paper, we elucidate the inference space of multimodal pLMs. Across three base models on four tasks, **vanilla sampling**, **reward-free guidance**, and **reward-guided search** act in complementary ways and together yield consistent, clear-margin gains over each model's default configuration, bringing multimodal pLMs close to, or even surpass, task-specific specialist models. Two unifying patterns emerge along the way: (1) inference preferences are largely task-oriented; and (2) the exploration-exploitation trade-off can be navigated bottom-up—first centering the foundational sampling distribution, then steering its logits toward the condition, and finally selecting among trajectories with a global quality signal.

Limitations

This work studies inference-time strategies for fixed multimodal pLMs on controlled benchmark tasks. Our goal is to improve the final-sample quality through vanilla sampling, classifier-free guidance, and reward-guided search, using the model's internal logits and quality signals. Although these results are useful across the tasks studied here, they do not directly translate to broader notions of biological utility or experimental success. We do not study richer deployment settings in which multimodal pLMs interact with external predictive models, human experts, laboratory feedback, or multi-stage protein design pipelines. These remain important directions for future work.

Besides, our analysis of performance-efficiency trade-offs remains at a very basic stage. Since the current work does not take computational efficiency as the optimization objective, the reported FLOPs and runtime objectively reflect the computational efficiency of inference protocols optimized purely for performance. We consider the multi-objective optimization of inference efficiency and benchmark performance under limited computational resources as future work.

Ethical Considerations

The primary societal impact of this work is to support scientific research and beneficial clinical and biomedical applications. However, as with any generative technology in biology, there is a theoretical risk of misuse in designing harmful biomolecules, such as pathogenic proteins. We believe this risk is mitigated by the fact that our study investigates the inference procedures of models rather than introducing new biological capabilities, operates on established computational benchmarks, and does not autonomously propose or experimentally validate hazardous biological designs. All data used in this work were obtained from publicly available sources, and no ethical approval was required.

Acknowledgments

The research described in this paper has been partially supported by the Key Special Project of National Natural Science Foundation of China (Project No. 72442017), the General Research Funds from the Hong Kong Research Grants Council (project No. PolyU 15200023, 15206024, and 15224524), and Internal research funds from Hong Kong Polytechnic University (project no. P0059586 and P0048625) and collaboration funds (project no. P0059105). This work was supported by computational resources provided by The Centre for Large AI Models (CLAIM) of The Hong Kong Polytechnic University.

References

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M Watkins, and 15 others. 2024. Openfold: retraining alphafold2 yields new insights

into its learning mechanisms and capacity for generalization. *Nature methods*, 21(8):1514–1524.

Christian B. Anfinsen. 1973. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. 2024. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*.

Tianlang Chen, Minkai Xu, Jure Leskovec, and Stefano Ermon. 2025. Rfg: Test-time scaling for diffusion large language model reasoning with reward-free guidance. *Preprint*, arXiv:2509.25604.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, and 3 others. 2022. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56.

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. 2026. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *International Conference on Learning Representations*, volume 2026, pages 61056–61109.

Wenqi Fan, Yi Zhou, Shijie Wang, Yuyao Yan, Hui Liu, Qian Zhao, Le Song, and Qing Li. 2025. Computational protein science in the era of large language models (llms). *Preprint*, arXiv:2501.10282.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. 2026. La-proteina: Atomistic protein generation via partially latent flow matching. In *International Conference on Learning Representations*, volume 2026, pages 67350–67395.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, and 6 others. 2025. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *Preprint*, arXiv:2207.12598.

Cheng-Yen Hsieh, Xinyou Wang, Daiheng Zhang, Dongyu Xue, Fei Ye, Shujian Huang, Zaixiang Zheng, and Quanquan Gu. 2025. Elucidating the design space of multimodal protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 24156–24175. PMLR.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. 2022. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8946–8970. PMLR.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and 15 others. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. 2024. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, volume 37, pages 52996–53021. Curran Associates, Inc.

Patrick Kunzmann, Tom David Müller, Maximilian Greil, Jan Hendrik Krumbach, Jacob Marcel Anter, Daniel Bauer, Faisal Islam, and Kay Hamacher. 2023. Biotite: new tools for a versatile python bioinformatics library. *BMC bioinformatics*, 24(1):236.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. 2025. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 95507–95545. Curran Associates, Inc.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Mengdi Liu, Xiaoxue Cheng, Zhangyang Gao, Hong Chang, Cheng Tan, Shiguang Shan, and Xilin Chen. 2025a. Protinvtree: Deliberate protein inverse folding with reward-guided tree search. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 28337–28360. Curran Associates, Inc.

Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, and Li Qing. 2025b. GLProtein: Global-and-local structure aware protein representation learning. In *Findings of the*

Association for Computational Linguistics: EMNLP 2025, pages 4355–4372, Suzhou, China. Association for Computational Linguistics.

Yunqing LIU, Yi Zhou, and Wenqi Fan. 2026. Enhancing molecular property predictions by learning from bond modelling and interactions. In *International Conference on Learning Representations*, volume 2026, pages 141589–141617.

Yunqing Liu, Yi Zhou, and Wenqi Fan. 2026. Geometric flow matching for molecular conformation generation via manifold decomposition. *Preprint*, arXiv:2605.25577.

Stephen Zhewen Lu, Jiarui Lu, Hongyu Guo, and Jian Tang. 2025. Towards protein sequence & structure co-design with multi-modal language models. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. 2025. Scaling up masked diffusion models on text. In *International Conference on Learning Representations*, volume 2025, pages 82974–82997.

Pascal Notin, Nathan Rollins, Yarin Gal, Chris Sander, and Debora Marks. 2024. Machine learning for functional protein design. *Nature biotechnology*, 42(2):216–228.

Yair Schiff, Subham Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. 2025. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, volume 2025, pages 43776–43821.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L M Gilchrist, Johannes Söding, and Martin Steinegger. 2024. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246.

Xinyou Wang, Zaixiang Zheng, Fei YE, Dongyu Xue, Shujian Huang, and Quanquan Gu. 2025. Dplm-2: A multimodal diffusion protein language model. In *International Conference on Learning Representations*, volume 2025, pages 35463–35490.

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, Basile I M Wicky, Nikita Hanikel, Samuel J Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, and 9 others. 2023. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100.

Jiaqi Wei, Xiang Zhang, Yuejin Yang, Wenxuan Huang, Juntai Cao, Sheng Xu, Xiang Zhuang, Zhangyang Gao, Muhammad Abdul-Mageed, Laks V. S. Lakshmanan, Chenyu You, Wanli Ouyang, and Siqi Sun. 2025. Unifying tree search algorithm and reward design for llm reasoning: A survey. *Preprint*, arXiv:2510.09988.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Jason Yim, Andrew Campbell, Emile Mathieu, Andrew Y. K. Foong, Michael Gastegger, Jose Jimenez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Frank Noe, Regina Barzilay, and Tommi Jaakkola. 2024. Improved motif-scaffolding with SE(3) flow matching. *Transactions on Machine Learning Research*.

Jason Yim, Marouane Jaakik, Ge Liu, Jacob Gershon, Karsten Kreis, David Baker, Regina Barzilay, and Tommi Jaakkola. 2025. Hierarchical protein backbone generation with latent and structure diffusion. *Preprint*, arXiv:2504.09374.

Jiahao Zhang, ZEQING ZHANG, Di Wang, and Lijie Hu. 2026. Controlling repetition in protein language models. In *International Conference on Learning Representations*, volume 2026, pages 80283–80331.

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

Zhuoqi Zheng, Bo Zhang, Kieran Didi, Kevin K. Yang, Jason Yim, Joseph L. Watson, Hai-Feng Chen, and Brian L. Trippe. 2025. Motifbench: A standardized protein design benchmark for motif-scaffolding problems. *Preprint*, arXiv:2502.12479.

Yi Zhou, Haohao Qu, Yunqing Liu, Shanru Lin, Le Song, and Wenqi Fan. 2026. Hd-prot: A protein language model for joint sequence-structure modeling with continuous structure tokens. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, page 13077–13088, New York, NY, USA. Association for Computing Machinery.

A Appendix

Table of Appendix:

- A.1. Supplementary Preliminaries
  - A.1.1. Protein Structure Tokenization
  - A.1.2. Base Model Descriptions
  - A.1.3. Evaluation Pipelines
- A.2. Details in Vanilla Sampling
  - A.2.1. Default Sampling Strategies
  - A.2.2. Optimal Vanilla Sampling Strategies
  - A.2.3. Ablation Study
- A.3. Details in Reward-free Guidance
  - A.3.1. Hyperparameter Selection
  - A.3.2. PDB Alignment Analysis on Unconditional Co-Generation Proteins.
  - A.3.3. Ablation Study on Motif-Scaffolding CFG
  - A.3.4. A Detailed Protein Structure Prediction Observation
- A.4. Details in Reward-guided Search
  - A.4.1. Implementation of Specialist Models
  - A.4.2. Beam Search for Multimodal pLMs
  - A.4.3. Hyperparameter Selection
  - A.4.4. Beam Search: Explore-then-Select
- A.5. Inference Efficiency Analysis

A.1 Supplementary Preliminaries

A.1.1 Protein Structure Tokenization

For a protein with L residues, its sequence is formulated as $\boldsymbol{s} = (s_1, s_2, \ldots, s_L)$, where each $s_i\,(1 \leq i \leq L)$ is a categorical variable denotes the identity of the i-th residue, generally involved in 20 standard amino acids $\mathbb{S}^{20} = \{\mathtt{A}, \mathtt{R}, \ldots, \mathtt{V}\}$. Meanwhile, the protein structure is firstly formulated by $\boldsymbol{x} = (x_1, x_2, \ldots, x_L)$, where $x_i \in \mathbb{R}^{n_i \times 3}$ is the Cartesian coordinates of the i-th residue's atoms.

To accommodate the discrete nature of language models, multimodal pLMs (Hayes et al., 2025; Wang et al., 2025; Hsieh et al., 2025) generally handle protein structures using quantization-based tokenizers. Similar to image tokenization, the protein structure tokenization process can be summarized under a VQ-VAE (van den Oord et al., 2017) framework:

$$\boldsymbol{x} \xrightarrow{\text{encoder \& quantizer}} \boldsymbol{z} \xrightarrow{\text{decoder}} \hat{\boldsymbol{x}},$$

where an encoder and a quantizer convert protein structure coordinates $\boldsymbol{x}$ into L discrete tokens $\boldsymbol{z} = (z_1, \ldots, z_L)$ each within a finite-size codebook, and the decoder reconstructs 3D coordinates $\hat{\boldsymbol{x}}$. Notably, following the practice of AlphaFold (Jumper et al., 2021) and ESMFold (Lin et al., 2023), the decoder of ESM3's structure tokenizer includes dedicated heads to compute structural quality scores, typically the predicted template modeling (pTM) score.

A.1.2 Base Model Descriptions

DPLM-2 (Wang et al., 2025) is a multimodal protein language model that extends a sequence-only pLM to model both sequence and structure. To incorporate structural information, it converts 3D coordinates $\boldsymbol{x}$ into discrete tokens $\boldsymbol{z}$ using a lookup-free quantization tokenizer[1]. Trained on high-quality protein data, DPLM-2 learns the joint distribution of protein sequence and structure, together with their marginals and conditionals. In our implementation, we use the pretrained 650M checkpoint[2], following the official instructions[3].

DPLM-2.1 (Hsieh et al., 2025) likewise models the joint distribution of protein sequence and structure. For structural modeling, it uses the same LFQ tokenizer as DPLM-2, with a vocabulary size of $2^{13} = 8192$, but predicts the 13 binary bits of each token rather than the corresponding 8192-way index. This finer-grained supervision should improve protein structure modeling. In our implementation, we use the pretrained 650M checkpoint[4], following the official instructions[5].

ESM3 (Hayes et al., 2025) is a multimodal generative language model that reasons over the sequence, structure, and function of proteins. Besides tokenized sequence and structure tracks, it also supports four additional parallel tracks, such as secondary structure labels, and can take structure coordinates as input. In this work, we focus on the sequence and structure tracks and enable coordinate input to improve protein structure modeling. In our implementation, we use the open-source ESM3-Open (1.4B) checkpoint[6], following the official codebase[7].

For notational uniformity, we denote the protein structure by discrete tokens $\boldsymbol{z}$ throughout the main text. In practice, however, we follow each

[1] https://huggingface.co/airkingbd/struct_tokenizer
[2] https://huggingface.co/airkingbd/dplm2_650m
[3] https://github.com/bytedance/dplm
[4] https://huggingface.co/airkingbd/dplm2_bit_650m
[5] https://github.com/bytedance/dplm
[6] https://huggingface.co/EvolutionaryScale/esm3-sm-open-v1
[7] https://github.com/evolutionaryscale/esm

model's recommended input format: ESM3 takes raw backbone coordinates $\boldsymbol{x}$ and motif coordinates $\boldsymbol{x}_{\text{motif}}$ for inverse folding and motif scaffolding, whereas DPLM-2 and DPLM-2.1 require the tokenized forms $\boldsymbol{z}$ and $\boldsymbol{z}_{\text{motif}}$. This input choice does not affect the generation formulation in Equation 1.

In our practice, all experiments were conducted on a single NVIDIA H20 (96GB) GPU. To compare the computational efficiency of different inference protocols, we also report the floating-point operations per second (FLOPs) and wall-clock runtime for each experiment.

A.1.3 Evaluation Pipelines

Unconditional protein sequence-structure co-generation jointly produces protein sequence and structure with only chain length specified (Geffner et al., 2026; Zhou et al., 2026). Following established protocols (Wang et al., 2025; Geffner et al., 2026), we sample 100 proteins at each target length (100, 200, 300, 400, and 500 residues), and report designability and diversity over both the designable subset and the full sample set. For each generated sample, we refold the sequence with ESMFold (Lin et al., 2023), where pLDDT is ESMFold's prediction confidence, and scRMSD and scTM measure the consistency between the ESMFold-predicted and the generated structures. A sample is classified as designable if scRMSD < 2.0 Å. On the designable subset, designability is summarized by the number of designable samples and diversity by the number of Foldseek (van Kempen et al., 2024) clusters of these samples. On the full sample set, designability is summarized by pLDDT and scTM, and diversity by the number of Foldseek clusters of all generated samples. Additionally, the secondary structure proportions are also calculated.

Motif scaffolding aims to generate protein scaffold structures that correctly embed specified target motifs (Geffner et al., 2026; Watson et al., 2023). Following established protocols (Wang et al., 2025; Yim et al., 2024), we generate 100 candidate scaffolds for each of the 24 benchmark problems, with scaffold length and motif order determined according to the specifications. For each generated candidate, we refold the sequence with ESMFold (Lin et al., 2023) and obtain scRMSD between the ESMFold-predicted structure and the generated structure, while motif-RMSD is computed directly between the motif residues of the generated structure and those of the ground-truth motif. A candidate is counted as successful if and only if it satisfies both global designability (scRMSD < 2.0 Å) and local motif preservation (motif-RMSD < 1.0 Å) (Zheng et al., 2025). The diversity of successful designs is quantified via Foldseek clustering (van Kempen et al., 2024).

Protein structure prediction infers a protein's 3D structure from its amino acid sequence (Jumper et al., 2021; Lin et al., 2023). Following established protocols (Wang et al., 2025; Hsieh et al., 2025), we evaluate on CAMEO 2022 and the PDB Date Split (Campbell et al., 2024), and quantify prediction accuracy by comparing the generated structures against their ground-truth natural structures with RMSD and TM-score.

Inverse folding generates amino acid sequences compatible with a specified target structure (Dauparas et al., 2022; Hsu et al., 2022). Following established protocols (Wang et al., 2025; Hsieh et al., 2025), we evaluate on CAMEO 2022 and the PDB Date Split. To assess structural self-consistency, we refold each generated sequence with ESMFold (Lin et al., 2023) and compute scTM relative to the target structure. As multiple sequences can yield about the same backbone structure, the amino acid recovery rate (AAR) is not a direct quality metric for inverse folding. Nevertheless, we report it to distinguish structural consistency from native-sequence recovery.

In our experiments, if no special indication is made, **every reported number is aggregated across five independent runs with distinct random seeds**, formulated as "mean ± std" or "mean (min, max)". The RMSD and TM-score are calculated using standard functions in OpenFold (Ahdritz et al., 2024) and TM-Tools (Zhang and Skolnick, 2005). The secondary structure is annotated via Biotite (Kunzmann et al., 2023). The number of clusters is obtained by clustering the generated structures via Foldseek (van Kempen et al., 2024), using this command:

```
foldseek easy-cluster ⟨input_path⟩
⟨output_path⟩ ⟨tmp_path⟩
-alignment-type 1 -cov-mode 0
-min-seq-id 0 -tmscore-threshold 0.5.
```

A.2 Details in Vanilla Sampling

A.2.1 Default Sampling Strategies

We document the default sampling strategies for each base model across the foundational tasks considered in this work. For DPLM-2 (Wang et al., 2025) and DPLM-2.1 (Hsieh et al., 2025), we fol-

Table 4: Default Sampling Strategies for Unconditional Co-Generation

	DPLM-2	DPLM-2.1	ESM3	ESM3
Multimodal Order	Synchronous	Synchronous	Seq $\rightarrow$ Struct	SS $\rightarrow$ Struct $\rightarrow$ Seq
Sampling Step T	500	500	L; 1	L
Unmasking	Stochastic	Stochastic	Random	Random
Remasking	Enabled	Enabled	Disabled	Disabled
Temp. & Annealing	From 2.0 to 0.1	From 1.1 to 0.1	From 1.0 to 0.0; 0.0	0.7

Table 5: Default Sampling for Motif-Scaffolding

	DPLM-2	DPLM-2.1	ESM3
Multimodal Order	Synchronous	Synchronous	Seq $\rightarrow$ Struct
Sampling Step T	500	500	$L/2$; $L/8$
Unmasking	Stochastic	Stochastic	Random
Remasking	Enabled	Enabled	Disabled
Temp. & Annealing	From 2.0 to 1.0	From 1.1 to 0.1	From 1.0 to 0.0

Table 6: Default Sampling for Structure Prediction

	DPLM-2	DPLM-2.1	ESM3
Sampling Step T	100	100	1
Unmasking	Deterministic	Deterministic	Deterministic
Remasking	Enabled	Enabled	Disabled
Temp. & Annealing	0.0	0.0	0.0

Table 7: Default Sampling for Inverse Folding

	DPLM-2	DPLM-2.1	ESM3
Sampling Steps T	100	100	8
Unmasking	Deterministic	Deterministic	Random
Remasking	Enabled	Enabled	Disabled
Temp. & Annealing	0.0	0.0	From 1.0 to 0.0

Table 8: Optimal Vanilla Sampling for Uncond. Co-Gen.

	DPLM-2	DPLM-2.1	ESM3
Multimodal Order	Synchronous	Synchronous	Synchronous
Sampling Step T	500	500	L
Unmasking	Stochastic	Stochastic	Stochastic
Remasking	Enabled	Enabled	Enabled
Temp. & Annealing	From 2.0 to 0.1	From 1.1 to 0.1	From 1.0 to 0.0

Table 9: Optimal Vanilla Sampling for Motif-Scaffolding

	DPLM-2	DPLM-2.1	ESM3
Multimodal Order	Synchronous	Synchronous	Synchronous
Sampling Step T	L	500	L
Unmasking	Stochastic	Stochastic	Random
Remasking	Enabled	Enabled	Enabled
Temp. & Annealing	From 0.5 to 0.0	From 0.5 to 0.0	From 0.7 to 0.0

Table 10: Optimal Vanilla Sampling for Structure Prediction

	DPLM-2	DPLM-2.1	ESM3
Sampling Step T	$L/8$	8	8
Unmasking	Random	Deterministic	Stochastic
Remasking	Enabled	Enabled	Enabled
Temp. & Annealing	From 0.1 to 0.0	From 0.1 to 0.0	From 0.5 to 0.0

low the official guideline and default hyperparameters[8]. For ESM3 (Hayes et al., 2025), we adhere to the implementation details described in the original appendix, official codebase[9], as well as established community practices (Yim et al., 2025; Zhou et al., 2026). Tables 4-7 present the default sampling configurations that serve as our primary baseline. These include the total sampling step, unmasking strategy, remasking allowance, sampling temperature, temperature annealing, and the multimodal sampling order.

Examining the default configurations of DPLM-2/-2.1 and ESM3 across four foundational tasks reveals two implicit heuristics: (1) the required diffusion steps follow the order unconditional co-generation $\geq$ motif scaffolding $>$ structure prediction $\approx$ inverse folding; and (2) the randomness during sampling scales as unconditional co-generation $\approx$ motif scaffolding $>$ structure prediction $\approx$ inverse folding. We aim to examine these task-dependent assumptions and derive more nuanced insights. Furthermore, the inference protocols of DPLM-2/-2.1 and ESM3 exhibit significant implementation mismatches. The original ESM3 codebase lacks native support for stochastic unmasking, remasking, and synchronous sequence-structure co-sampling. Resolve this, we hope to standardize the ESM3 inference protocol and rigorously evaluate the effectiveness of each vanilla sampling strategy under a unified framework.

In addition to these, we retain several default configurations after careful consideration. First, to preserve each model's inherent training-inference alignment, we maintain the original diffusion schedules (linear for DPLM-2 and DPLM-2.1, cosine for ESM3) and temperature annealing functions (linear for DPLM-2 and DPLM-2.1, quadratic for ESM3). Second, for unmasking, DPLM-2 and DPLM-2.1 evaluate positional confidence using log-probabilities, whereas ESM3 relies on entropy, which has been confirmed in prior work to have little impact (Lu et al., 2025). Although these settings introduce minor methodological differences, they should not detract from our core investigation into the exploration-exploitation tradeoff.

A.2.2 Optimal Vanilla Sampling Strategies

Tables 8-11 present the optimal vanilla sampling configurations we identified. Configurations that deviate from the defaults are highlighted in blue.

[8] https://github.com/bytedance/dplm

[9] https://github.com/evolutionaryscale/esm

Table 11: Optimal Vanilla Sampling for Inverse Folding

	DPLM-2	DPLM-2.1	ESM3
Sampling Step T	1	8	8
Unmasking	Random	Stochastic	Deterministic
Remasking	Enabled	Enabled	Enabled
Temp. & Annealing	From 0.1 to 0.0	From 0.1 to 0.0	From 0.5 to 0.0

A.2.3 Ablation Study

Due to space constraints, we present the comprehensive vanilla sampling ablation results across models and tasks in Figure 5. Each panel reports task-specific metrics across varying sampling hyperparameters for a single base model, with the default-setting performance marked for reference. The caption explains how to read the figure.

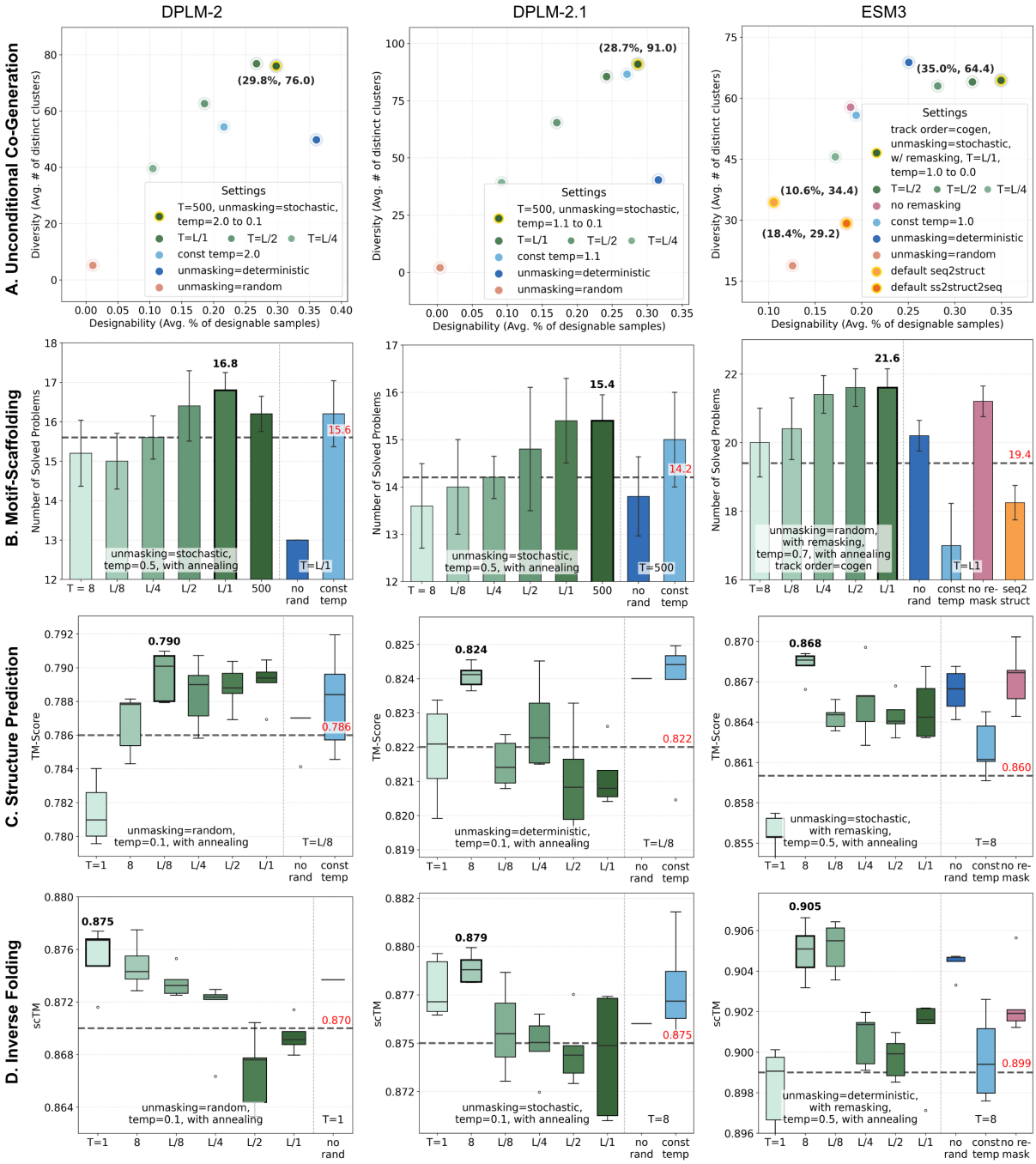

Figure 5: **Vanilla sampling: optimal configurations and ablation results**. **A.** Pareto comparison between designability and diversity in unconditional protein sequence-structure co-generation. The first legend shows the optimal configuration, and each subsequent legend changes one setting at a time. **B-D.** The green bars to the left of the vertical dotted line show ablations over the total sampling step T, with the other vanilla sampling settings listed. The remaining bars show one-at-a-time changes to the optimal configuration: "no rand" denotes deterministic unmasking with temperature $= 0.0$ (argmax sampling); "const temp" and "no remasking" denote disabled temperature annealing and remasking, respectively.

A.3 Details in Reward-free Guidance

A.3.1 Hyperparameter Selection

To select the CFG scales, we have conducted a limited hyperparameter search for each base model across the four tasks. Table 12 lists the selected scales, with those yielding gains over the optimal vanilla sampling underlined, and the corresponding results are reported in the main text.

A notable observation from this hyperparameter search concerns the sequence-structure asymmetry in multimodal pLMs. Since the two tracks are inherently imbalanced in tokenization, some differences in behavior are expected. However, the asymmetry in CFG response is substantially more pronounced for DPLM-2/-2.1 than for ESM3, as reflected in the contrast between protein structure prediction and inverse folding in Figure 6. For DPLM-2 and DPLM-2.1, CFG provides no benefit for protein structure generation but yields clear gains when guiding inverse folding with a selected guidance scale. By contrast, ESM3 can positively respond to CFG in both directions, behaving more balanced across the two tasks.

A plausible explanation is that DPLM-2 and DPLM-2.1 are developed through modality extension, from sequence-only to multimodal pLMs, so the two tracks may not be internalized equally well. ESM3, trained natively at a larger scale on multimodal protein data, appears to internalize sequence and structure conditions more evenly, leading to a more balanced response to guidance across the two tracks. This is consistent with our insight that the exploitation effect of CFG depends on how well the base model has internalized the corresponding conditional signal.

A.3.2 PDB Alignment Analysis on Unconditional Co-Generation Proteins

To further confirm the effect of CFG in making ESM3's unconditional co-generation more "protein-like", we align the generated samples with PDB proteins. Specifically, we search each generated sample against the PDB database using Foldseek (van Kempen et al., 2024) and recorded the highest TM-score from the alignment against all PDB proteins, denoted as PDB-TM.

In our benchmark of 500 generated proteins, ESM3 in the optimal vanilla sampling configuration yields an average of 387.2 samples that could be aligned to PDB proteins, with a mean PDB-TM of 0.548 ± 0.316. In contrast, ESM3 with cross-modal CFG yields an average of 473.0 samples with aligned PDB proteins, with a mean PDB-TM of 0.821 ± 0.156. These results demonstrate that cross-modal CFG makes ESM3's generations structurally closer to PDB proteins.

The concrete Foldseek command is:

```
foldseek easy-search ⟨input_path⟩
⟨database_path⟩ ⟨output_path⟩
⟨tmp_path⟩ -exhaustive-search
-alignment-type 1 -tmscore-threshold 0.0
-format-output query,target,qtmscore.
```

A.3.3 Ablation Study on Motif-Scaffolding CFG

To validate our dual-condition CFG design for motif scaffolding, we compare it against two single-condition variants on ESM3. In all cases, the conditional logits remain those under full conditioning, namely $\mathbf{l}^{s,(t)}_{\text{cond}}, \mathbf{l}^{z,(t)}_{\text{cond}} = f_\theta(\boldsymbol{s}^{(t)}, \boldsymbol{z}^{(t)})$. For motif-only CFG, the unconditional branch drops only the motif condition while retaining cross-modal context, yielding

$$\mathbf{l}^{s,(t)}_{\text{uncond}}, \mathbf{l}^{z,(t)}_{\text{uncond}} = f_\theta\big(\boldsymbol{s}^{(t)} \setminus \boldsymbol{s}_{\text{motif}}, \boldsymbol{z}^{(t)} \setminus \boldsymbol{z}_{\text{motif}}\big).$$

For cross-modal CFG, the unconditional branch drops only the opposite-track context while retaining motif information, yielding

$$\mathbf{l}^{s,(t)}_{\text{uncond}} = f_\theta\big(\boldsymbol{s}^{(t)}, \varnothing\big), \quad \mathbf{l}^{z,(t)}_{\text{uncond}} = f_\theta\big(\varnothing, \boldsymbol{z}^{(t)}\big).$$

Table 13 reports the ablation results. Both single-conditioning CFG variants improve over the optimal vanilla sampling, while the full dual-conditioning CFG achieves the strongest gains. These results show that explicit motif guidance and implicit cross-modal guidance are both useful and are most effective when combined.

A.3.4 A Detailed Protein Structure Prediction Observation

As noted in Table 3, ESM3 exhibits an interesting divergence between RMSD and TM-score with sampling strategies adjusted. Moving from default sampling ($T = 1$, deterministic unmasking, no remasking, temperature $= 0.0$) to the optimized vanilla sampling ($T = 8$, stochastic unmasking, remasking, temperature annealing from 0.5 to 0.0) degrades RMSD while also improving TM-score on both datasets. Applying CFG on top of the selected vanilla sampling strategy mitigates this RMSD–TM-score divergence, improving both metrics simultaneously.

Table 12: Selected CFG scales.

	Uncon. Co-Gen.	Motif-Scaffolding	Structure Prediction	Inverse Folding
DPLM-2	$w_s = 2.0, w_z = 1.5$	$w_s = 1.5, w_z = 1.0$	$w_z = 1.5$	$w_s = 4.0$
DPLM-2.1	$w_s = 2.0, w_z = 1.5$	$w_s = 1.5, w_z = 1.0$	$w_z = 1.5$	$w_s = 2.0$
ESM3	$w_s = 2.0, w_z = 3.0$	$w_s = 2.0, w_z = 2.0$	$w_z = 2.0$	$w_s = 2.0$

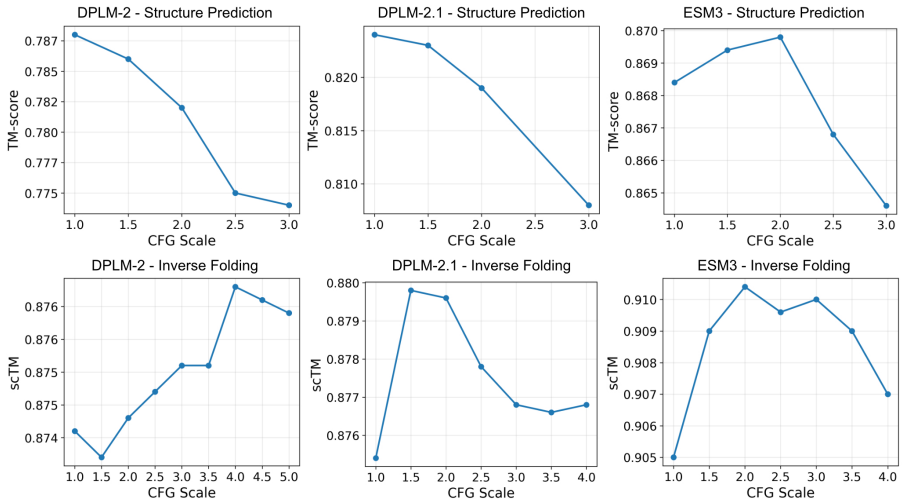

Figure 6: CFG scale analysis for protein structure prediction and inverse folding on CAMEO2022.

Table 13: Ablation study on motif-scaffolding CFG.

	#Solved / 24	Success Rate	#Clusters
ESM3 $_{\text{Vanilla}}$	21.6 (21, 22)	30.9 ± 1.1 %	110.0 (100, 124)
ESM3 $_{\text{CFG (Motif-only)}}$	22.6 (22, 23)	35.3 ± 0.6 %	146.8 (127, 158)
ESM3 $_{\text{CFG (Cross-Modal)}}$	22.4 (21, 23)	36.6 ± 1.4 %	155.0 (144, 166)
ESM3 $_{\text{CFG (Both)}}$	22.8 (22, 23)	37.5 ± 1.0 %	179.4 (165, 198)

Figure 7 provides two case studies to illustrate this behavior. For some challenging targets, default single-pass argmax decoding yields low-quality predictions with incorrect secondary-structure organization. Multi-step iterative sampling recovers more plausible global topologies, which improves TM-score, but local fragment misorientations can persist and keep RMSD high. CFG further refines these local geometries, reducing this mismatch and yielding predictions with both more native-like global folds and more accurate local alignments.

A.4 Details in Reward-guided Search

A.4.1 Implementation of Specialist Models

For unconditional protein sequence-structure co-generation, we run La-Proteina (Geffner et al., 2026) using the pre-trained checkpoint[10] and follow the official guideline[11], using the default noise scales of 0.1 for the alpha-carbon atoms and 0.1 for the latent variables. We also benchmark the official SVDD implementation in ESM3 following the official tutorial[12]. Given a desired length L, it employs a sequence-to-structure sampling order: sequence tokens are first sampled over $L/8$ steps with random unmasking, no remasking, and temperature annealing from 1.0 to 0.0, after which structure tokens are predicted in a single argmax step. On top of this default procedure, SVDD branches 10 candidates at each step and retains the one with the best pTM. For motif scaffolding, we run La-Proteina using the motif-scaffolding checkpoint[13] with the default sampling settings. For inverse folding, we run ProteinMPNN (Dauparas et al., 2022) using the official checkpoint "v_48_020.pt" with temperature $= 0.1$.

A.4.2 Beam Search for Multimodal pLMs

In this study, we adapt beam search to the discrete diffusion sampling of multimodal pLMs, operating jointly over protein sequence and structure $(\boldsymbol{s}, \boldsymbol{z})$, as summarized in Algorithm 1. This procedure maintains N trajectories in parallel in the set $\mathcal{T}$.

At each step t, the search layer calls $\textsc{Expand}(\cdot)$, which applies the underlying sampler for one step independently to each retained trajectory. At general steps, $\textsc{Expand}(\mathcal{T}, 1)$ simply advances each retained trajectory once. Meanwhile, every K steps is an expansion step to periodically expand,

[10] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/resources/ld2_ucond_tri_512.ckpt/files

[11] https://github.com/NVIDIA-Digital-Bio/la-proteina

[12] https://github.com/evolutionaryscale/esm/blob/main/cookbook/tutorials/5_guided_generation.ipynb

[13] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/resources/ld4_motif_idx_aa.ckpt/files

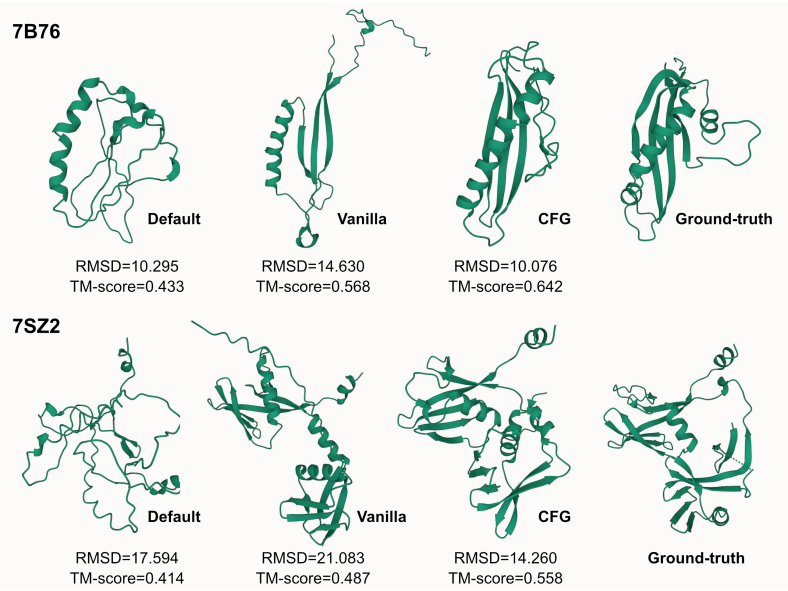

Figure 7: Two cases to illustrate ESM3's RMSD-TM-score divergence in protein structure prediction.

Algorithm 1 Generalized Beam Search

Require: initial state $(\boldsymbol{s}^{(T)}, \boldsymbol{z}^{(T)})$, total steps T, beam width N, branching factor B, scoring interval K, reward $r(\cdot)$

$\mathcal{T} \leftarrow \{(\boldsymbol{s}_n^{(T)}, \boldsymbol{z}_n^{(T)})\}_{n=1}^{N}$
for $t = T, T{-}1, \ldots, 1$ **do**
  if $t \bmod K = 0$ **and** $t < T$ **then**
    $\mathcal{C} \leftarrow \textsc{Expand}(\mathcal{T}, B)$
    $\hat{r}(\boldsymbol{c}) \leftarrow r(\textsc{QuickUnmask}(\boldsymbol{c}))$, $\forall \boldsymbol{c} \in \mathcal{C}$
    $\mathcal{T} \leftarrow \textsc{Select}(\mathcal{C}, \hat{r}, N)$
  else
    $\mathcal{T} \leftarrow \textsc{Expand}(\mathcal{T}, 1)$
  end if
end for
return $\textsc{Select}(\mathcal{T}, r, 1)$

score, and prune trajectories: First, $\textsc{Expand}(\mathcal{T}, B)$ yields $N \times B$ candidate trajectories, collected in set $\mathcal{C}$. Second, since partially masked candidates cannot be directly evaluated, we apply a single-forward argmax decoding $\textsc{QuickUnmask}(\cdot)$ to predict all remaining masked tokens and compute $r(\cdot)$ on the resulting proteins, leading to the estimated rewards $\hat{r}$. Third, the selection rule $\textsc{Select}(\cdot)$ then chooses N beams from $\mathcal{C}$ according to $\hat{r}$ for the next step of denoising.

After all T steps, the final output is produced by applying the same $\textsc{Select}(\cdot)$ rule to the N completed beams with target width 1. Notably, this formulation unifies two typical cases: $N{=}1, K{=}1$ recovers single-trajectory search with per-step reranking, i.e., SVDD (Li et al., 2025), and $B{=}1, K{=}T$ reduces to Best-of-N sampling.

Within this framework, the task-aware design choices in protein modeling reduce to two questions: how to instantiate the reward $r(\cdot)$, and how to instantiate the selection rule $\textsc{Select}(\cdot)$.

- **Reward**. While external models such as ESMFold could in principle serve as the scorer, we rely on signals internal to the multimodal pLM. This reflects our choice to keep all conditioning signals internal to the model, both to preserve the "all-in-one" nature of multimodal pLMs and to ensure a fair comparison with vanilla and guided sampling. As described in Section A.1.1, ESM3's protein structure tokenizer natively produces pTM scores during de-tokenization, which we use as global, trajectory-level rewards. A candidate on the structure track is scored directly (named structural pTM), whereas a candidate on the sequence track is first folded with an additional single-pass argmax decoding and then scored (named foldability pTM).
- **Selection**. The choice of $\textsc{Select}(\cdot)$ is aligned with the exploration-exploitation profile of each task. For the exploitation-biased and balanced tasks, $\textsc{Select}(\cdot)$ simply returns the top-N (or top-1) candidates ranked by the corresponding pTM: structural pTM for structure prediction, foldability pTM for inverse folding, and their sum for motif-scaffolding. For the exploration-biased task (unconditional co-generation), we optionally use a threshold-based random selection that uniformly samples from candidates satisfying both per-track criteria (foldability pTM > 0.8 and structural pTM > 0.8), which preserves diversity while enforcing a quality lower bound. Notably, pTM > 0.8 is a commonly used thresh-

Table 14: ESM3 $_{\text{Beam}}$ hyperparameter selection: unconditional co-generation

		Designable Subset		All Samples			
N	B	#Design	#Clusters	pLDDT	scTM	#Clusters	α / β %
2	1	318.0	137.0	89.529	0.955	194.0	45.7 / 12.8
4	**1**	350.4	139.0	90.704	0.964	177.6	47.9 / 11.6
8	1	342.0	132.0	90.632	0.963	175.0	47.6 / 11.7
4	2	385.0	119.0	91.342	0.973	138.0	50.2 / 10.5
4	4	361.0	117.0	91.574	0.970	135.0	50.0 / 10.9
4	8	382.0	107.0	91.251	0.973	135	50.7 / 10.4

Table 15: ESM3 $_{\text{Beam}}$ hyperparameter selection: motif scaffolding

N	B	#Solved	Success Rate	#Clusters
1	2	23.0	38.2 %	174.0
1	4	23.0	38.0 %	186.0
2	1	23.0	43.7 %	222.0
2	2	23.0	46.0 %	240.0
2	4	23.0	47.8 %	233.0
4	**1**	23.0	49.9 %	261.0
4	2	23.0	52.9 %	228.5
4	4	23.0	54.0 %	221.0

old (Hayes et al., 2025).

A.4.3 Hyperparameter Selection

In the reward-guided beam search, to select the beam width N, branching factor B, and scoring interval K, we conduct a limited hyperparameter search and obtain some primary observations.

For unconditional cogeneration, we examine $N \in \{2, 4, 8\}$ and $B \in \{1, 2, 4, 8\}$ while fixing $K = T/5$. For motif scaffolding, we examine $N \in \{1, 2, 4\}$ and $B \in \{1, 2, 4\}$ while fixing $K = T/5$. The results are reported in Tables 14 and 15, respectively. In both tables, only the selected (gray-highlighted) rows are aggregated over five random seeds, whereas other rows report single-seed results. Overall, increasing N consistently improves designability, suggesting that a wider beam is more likely to preserve at least one high-quality trajectory for final selection. By contrast, increasing B can reduce diversity, especially in a wide beam. Although a larger branching factor expands intermediate step exploration, it also leads to more aggressive pruning, so trajectories that might later yield diverse outcomes can be discarded too early, causing the search to concentrate on trajectories that appear better at the current step.

For protein structure prediction and inverse folding, we examine $N \in \{1, 2, 4\}$ and $B \in \{1, 2, 4\}$ on the CAMEO 2022 dataset while fixing $K = 1$, and report the results. As Figure 8 shows, larger N and B consistently improve performance on both tasks. We attribute this trend to the fact that protein structure prediction and inverse folding are already

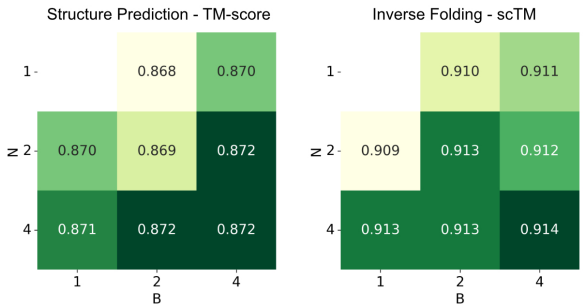

Figure 8: ESM3 $_{\text{Beam}}$ hyperparameter selection. Left: protein structure prediction; right: inverse folding.

Table 16: ESM3 $_{\text{Beam}}$ explore-then-select: unconditional protein sequence-structure co-generation

(a) Designable Subset

Temp.	#Design ↑	#Clusters ↑
From 1.0 to 0.0	393.8 (383, 405)	117.0 (107, 127)
1.0	350.4 (336, 360)	139.0 (132, 145)

(b) All Samples

Temp.	pLDDT ↑	scTM ↑	#Clusters ↑	α / β %
1.0 to 0.0	92.106 ± 4.781	0.974 ± 0.048	144.0 (137, 152)	54.7 / 10.1
1.0	90.704 ± 5.525	0.964 ± 0.056	177.6 (170, 186)	47.9 / 11.6

relatively exploitation-biased. A larger beam width is more likely to retain high-quality trajectories, while a larger branching factor makes it easier to identify the best trajectory during search.

A.4.4 Beam Search: Explore-then-Select

Across three of the four multimodal protein modeling tasks, beam search works better when the base sampler is made slightly more exploratory. For each task, we start from the selected vanilla sampling configuration (Table 8-11) and CFG scale (Table 12), and modify only one vanilla sampling setting. Table 16-18 report the ablation results.

For unconditional co-generation, replacing temperature annealing from 1.0 to 0.0 with a constant temperature of 1.0 trades designability for diversity: the number of designable samples drops from 393.8 to 350.4, while the number of clusters on the designable subset rises from 117.0 to 139.0. For motif scaffolding, increasing the initial temperature from 0.7 to 1.0 improves both success rate (48.9% to 49.9%) and diversity (214.6 to 261.0 clusters). For inverse folding, switching the unmasking strategy from deterministic to stochastic yields modest but consistent gains in scTM on both CAMEO 2022 and the PDB Date Split.

We suspect that, without this additional exploration, the beams remain too similar for reward-based selection to meaningfully distinguish among the protein generation candidates. When exploring

Table 17: ESM3 $_{\text{Beam}}$ explore-then-select: motif-scaffolding

Temperature	#Solved / 24	Success Rate	#Clusters
From 0.7 to 0.0	23.0 (23, 23)	48.9 ± 0.9 %	214.6 (202, 226)
From 1.0 to 0.0	23.0 (23, 23)	49.9 ± 0.7 %	261.0 (250, 278)

Table 18: ESM3 $_{\text{Beam}}$ explore-then-select: inverse folding

	CAMEO		PDB Date Split	
Umasking	scTM ↑	AAR	scTM ↑	AAR
Deterministic	0.910 ± 0.134	45.7	0.954 ± 0.075	49.1
Stochastic	0.914 ± 0.129	46.0	0.956 ± 0.069	49.3

a too-narrow candidate space, the search could fail to achieve substantial diversity gains in more open-ended, exploratory-biased tasks or self-consistency gains in exploitation-biased tasks.

A.5 Inference Efficiency Analysis

As shown in Tables 1-3, in addition to benchmark performance, we report the specific FLOPs and actual wall-clock runtime on a unified device (a single Nvidia H20 GPU). Based on these results, we can draw the following analysis.

FLOPs and Forward Passes. For benchmarks with a fixed number of tokens, FLOPs are roughly proportional to the number of forward passes. In the vanilla sampling stage, the number of passes equals the diffusion steps T. In the CFG stage, the model combines conditional and unconditional outputs, requiring 2 passes per step for structure prediction and inverse folding, and 3 passes for motif scaffolding and unconditional co-generation. That is, $2T$ and $3T$ passes in total, respectively.

In the reward-guided search stage, let N be the beam width, B the branching factor, and K the scoring interval. The number of branching (scoring) steps is $S = \lfloor (T-1)/K \rfloor$. The total forward passes equal $((T-S)N + SNB) \times$ per-step cost (2 or 3 passes depending on the CFG formulation), plus an additional $(SNB + N) \times$ reward cost (1 to 3 passes depending on the reward type). Specifically, our beam search inference requires $2((T-S)N + SNB) + (SNB + N)$ passes for structure prediction, $2((T-S)N + SNB) + 2(SNB + N)$ passes for inverse folding, and $3((T-S)N + SNB) + 3(SNB + N)$ passes for motif scaffolding and unconditional sequence-structure co-generation.

Runtime vs. FLOPs Scaling. Due to GPU parallelization optimizations, the actual inference runtime often grows sublinearly with FLOPs. Depending on the base model implementation, if the GPU utilization is close to 100 percent, the runtime scales nearly proportionally with FLOPs. Conversely, if there is spare GPU capacity, the runtime may grow slowly even when FLOPs increase significantly.

Suboptimal Efficiency in Default Inference. Default inference protocols are often suboptimal in both performance and efficiency. Through our first-stage vanilla sampling exploration, we can achieve simultaneous runtime reduction and performance improvement for DPLM-2/2.1 (structure prediction, inverse folding) and ESM3 (unconditional co-generation, inverse folding). This is achieved by reducing the total number of pLM forward passes, i.e., setting fewer diffusion steps and enabling synchronous sequence-structure co-sampling.

Performance-Efficiency Balance in CFG Inference. Our classifier-free guidance (CFG) strategies offer a superior performance-efficiency balance, delivering various degrees of performance improvement with only 1 to 3 times runtime overhead. Especially for ESM3, whose implementation leaves spare GPU capacity, resulting in only a moderate increase in runtime despite doubling the FLOPs, yielding practical benefits across all four tasks. In the unconditional co-generation benchmark, ESM3 runtime increases by about 1.46 times, while the number of designable clusters increases by about 1.81 times. In the motif scaffolding benchmark, ESM3 runtime remains nearly flat, while the number of clusters increases by a factor of 1.63. Similarly, for structure prediction and inverse folding, ESM3 performance improves while runtime remains nearly constant (<1.1 times).

Beam Search Efficiency. Table 1-3 results acknowledge that reward-guided beam search significantly increases inference costs. Meanwhile, as shown in the FLOPs estimation formats, increasing the beam width, branching factor, or scoring frequency multiplies the computational overhead. However, we believe this trade-off is justifiable for two reasons. First, our approach is training-free. Unlike specialized models or post-training studies that require multi-node clusters and weeks of training, we adjust sampling strategies using a single GPU for at most several hours. Second, the performance gains at this stage demonstrate that pLMs can compete with specialized state-of-the-art models, overturning prior consensus on multimodal pLMs such as ESM3.