DPLM-2: A MULTIMODAL DIFFUSION PROTEIN LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Proteins are essential macromolecules defined by their amino acid sequences, which determine their three-dimensional structures and, consequently, their functions in all living organisms. Therefore, generative protein modeling necessitates a multimodal approach to simultaneously model, understand, and generate both sequences and structures. However, existing methods typically use separate models for each modality, limiting their ability to capture the intricate relationships between sequence and structure. This results in suboptimal performance in tasks that requires joint understanding and generation of both modalities. In this paper, we introduce DPLM-2, a multimodal protein foundation model that extends discrete diffusion protein language model (DPLM) to accommodate both sequences and structures. To enable structural learning with the language model, 3D coordinates are converted to discrete tokens using a lookup-free quantization-based tokenizer. By training on both experimental and high-quality synthetic structures, DPLM-2 learns the joint distribution of sequence and structure, as well as their marginals and conditionals. We also implement an efficient warm-up strategy to exploit the connection between large-scale evolutionary data and structural inductive biases from pre-trained sequence-based protein language models. Empirical evaluation shows that DPLM-2 can simultaneously generate highly compatible amino acid sequences and their corresponding 3D structures eliminating the need for a two-stage generation approach. Moreover, DPLM-2 demonstrates competitive performance in various conditional generation tasks, including folding, inverse folding, and scaffolding with multimodal motif inputs.

1 INTRODUCTION

Proteins are macromolecules that execute crucial roles in every living organism. They are characterized by their amino acid sequences and three-dimensional structure, where the sequence determines the structure, which in turn governs the protein's function. Generative modeling for proteins has made significant strides in recent years. Among them, diffusion models (Ho et al., 2020; Song et al., 2020) exhibit great success in protein structure-based generative modeling (Watson et al., 2023; Yim et al., 2023). Meanwhile, large-scale protein language models (Rives et al., 2019; Lin et al., 2022), trained on evolutionary-scale sequence database, have become one of the most important cornerstones in sequence-based foundation models for protein sequence representation learning and generation. Remarkably, DPLM (Wang et al., 2024), a discrete diffusion (Austin et al., 2021) based protein language models, has exhibited the state-of-the-art performance in both sequence generation and understanding, addressing a wide range of sequence-oriented applications.

Many protein engineering applications, *e.g.*, motif-scaffolding (Watson et al., 2023; Yim et al., 2024) and antibody design (Jin et al., 2021; Kong et al., 2022; Zhou et al., 2024), require jointly determine both structure and sequence. However, the aforementioned approaches mostly employ generative models for one modality (either sequence or structure) and resort to separate models (Jumper et al., 2021; Dauparas et al., 2022) for the other. This highlights the pressing need for multimodal protein generative models that can integrate both sequence and structure, enabling a more comprehensive understanding of protein behaviors and functions. This, therefore, raises the following question:

Can we build a multimodal protein foundation model to simultaneously model, understand, and generate both sequences and structures?

To pursue this goal, Multiflow (Campbell et al., 2024) is a recent effort for structure-sequence co-generation that incorporates sequences into structure-based generative models using multimodal flow matching. Despite its impressive structure generation capability, Multiflow exhibits suboptimal

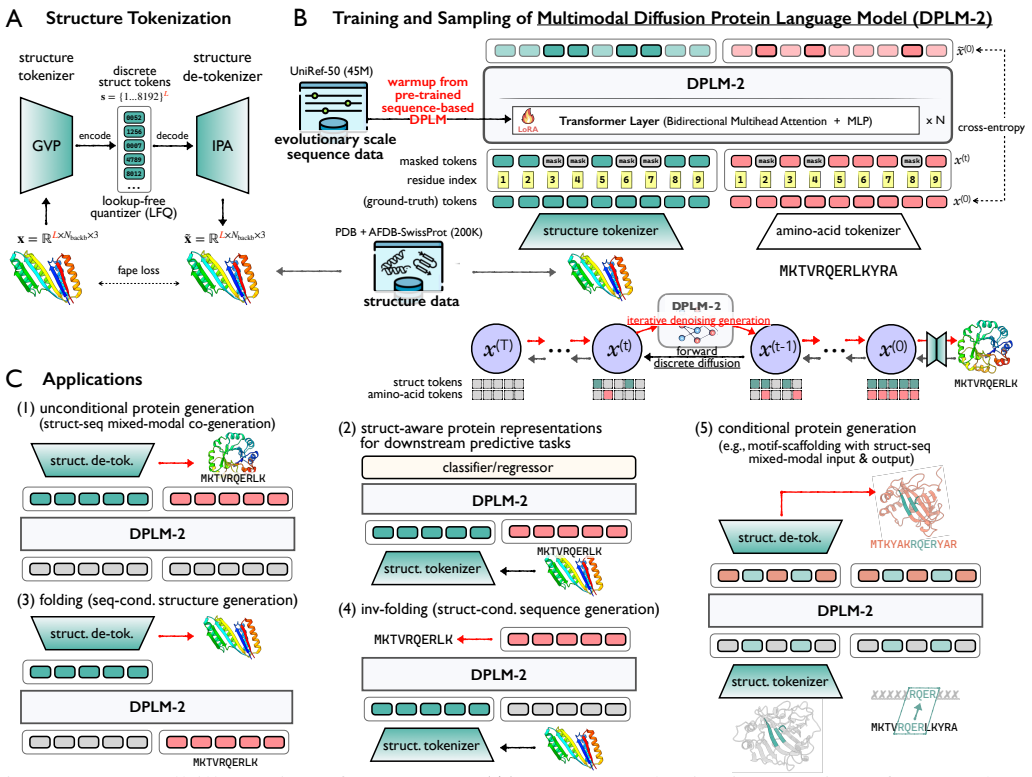


Figure 1: *Overall illustration of* DPLM-2. **(A)** structure tokenization consists of a GVP-based encoder to yield invariant backbone geometric features, a lookup-free quantizer (LFQ) to discretize encoded structural features into structure tokens within a codebook, and an IPA-based decoder as de-tokenizer to convert structure tokens back to backbone atomic coordinates. **(B)** multimodal learning and generation of protein structure and sequence with DPLM-2. **(C)** various applications of DPLM-2 as a protein foundation model: (1) unconditional protein sequence-structure mixed-modal co-generation; (2) protein sequence-structure joint representation for predictive tasks; (3) structure prediction; (4) fixed-backbone sequence generation; (5) conditional protein generation with structure-sequence mixed-modal input and output.

performance in co-generating structurally-compatible sequences and consequently resorts to instance-level knowledge distillation from ProteinMPNN (Dauparas et al., 2022). Furthermore, it completely falls short in protein folding for given sequences, showing Mulitflow's inadequacy in sequence understanding. We argue that this bottleneck arises from the absence (co-)evolutionary inductive bias derived from massive pre-training from sequence database, as prior studies have demonstrated that the evolutionarily-informed representations learned by pre-trained protein language models implicitly capture structural information enables direct structure prediction (Lin et al., 2022). As a consequence, the limitation in sequence understanding and generation renders Multiflow inadequate as a multimodal protein generative foundation.

Inspired by the connection between evolutionary knowledge and spatial interactions, we suggest that sequence-based generative language models like DPLM, with their strong sequence generation and predictive abilities, hold great promise as a foundation for multimodal learning for proteins. Despite its exciting potential, this approach presents two key challenges: (1) language models cannot directly handle continuous data like structure; and (2) language models heavily necessitate sufficient scale of data and compute resources while structure data is much smaller compared to sequence databases.

In this paper, we address the aforementioned questions by introducing DPLM-2, a multimodal protein foundation model that advances the state-of-the-art discrete diffusion-based protein language model (*i.e.*, DPLM) to accommodate both sequences and structures. By training on both experimental and high-quality synthetic structures, DPLM-2 learns the joint distribution of sequence and structure, as well as their marginals and conditionals. We present several key receipts to facilitate multimodal learning in DPLM-2: (1) the core difficulty lies in enabling the language model to learn structural information, which is challenging and remains elusive, for which we develop a lookup-free quantization (LFQ, Yu et al., 2023) structure tokenizer to convert 3D coordinates to discrete tokens and vice versa (Fig. 1A, §3.3); (2) we implement an efficient warm-up strategy to exploit the connection

between large-scale evolutionary data and structural inductive biases from pre-trained sequence-based DPLM (Fig. 1B, §3.2); and (3) we also address the exposure bias problem in discrete diffusion for sequence learning (Ranzato et al., 2016; Bengio et al., 2015) by a self-mixup training strategy that leads to enhanced generation quality and diversity.

We highlight our main contributions and findings as follows:

(i) We present DPLM-2, a multimodal protein generative language model that aims to simultaneously model, understand and generate protein structure and sequence. We show that it can be fairly efficient and effective to obtain a mulitmodal protein model with moderate amount of high-quality data, a decent structure tokenizer and publicly-accessible sequence-only pre-trained language models.

(ii) As a mulitmodal generative model, DPLM-2 enables unconditional co-generation of designable and diverse proteins that guarantees consistency between structure and sequence (Fig. 1C(1)). Our empirical evaluation show that DPLM-2 attains competitive co-generation performance compared to structure-based generative approaches, while DPLM-2's generated proteins better align with the characteristics of natural proteins regarding secondary structure statistics (§4.1).

(iii) In addition, DPLM-2 allows various conditional generation tasks by its multimodal nature, ranging from (sequence-conditioned) folding (Fig. 1C(3), §4.2), (structure-conditioned) inverse-folding (Fig. 1C(4), §4.3), to more successful motif-scaffolding given multimodal motif conditioning (Fig. 1C(5), §4.4).

(iv) Last but not least, we demonstrate that the structure-aware protein representation learned by DPLM-2 brings additional benefit for a range of protein predictive tasks (Fig. 1C(2), §4.5).

Concurrent work. During the development of DPLM-2, we became aware of the recently proposed multimodal generative protein language model, ESM3 (Hayes et al., 2024), which also jointly models tokenized structure and sequence using a generative masked language model. While both models aim for similar goals, DPLM-2 differs from ESM3 in several key aspects: **(1)** *Multimodal protein generation:* DPLM-2 treats structure and sequence modalities equally by design and emphasizes the simultaneous co-generation of compatible protein sequence and structure, whereas ESM3 is a sequence-first model (other modalities are subject to dropout during training) and generates in cascaded modality-by-modality manner. **(2)** *Data and compute efficiency:* ESM3 seeks to perform mulimodal pre-training from scratch using a huge amount of synthetic data, with modal size ranging from 1.4B to 98B. With strict license and absence of training infrastructure, this prohibits community from replicating for customized purposes. In contrast, DPLM-2 leverages much smaller datasets (PDB + SwissProt) and builds on open-source, pre-trained sequence-based DPLM (150M/650M/3B), which leverages DPLM's learned evolutionary knowledge and inherits strong sequence understanding and generation capabilities. We are also committed to open-source our models, training and inference code to democratize multimodal generative protein LM to benefit the community. Overall, we believe DPLM-2 provides unique contributions to the community.

2 Preliminaries

2.1 Generative Modeling for Protein

The aim of generative protein modeling is to estimate the underlying distribution $\text{prot} \sim q(\text{prot})$ of the protein data of our interest by learning a probabilistic model $p_\theta(\text{prot})$. Here $\text{prot} = (r_1, r_2, \dots, r_L)$ denotes a protein with L residues, where each residue $r_i = (s_i, x_i)$ is represented by two major modalities, *i.e.*, $s_i \in \{0,1\}^{|\mathcal{S}|}$ is a categorical variable for its amino acid type in $\mathcal{S} = \{1, ..., 20\}$, and $x_i \in \mathbb{R}^{N_{\text{atoms}} \times 3}$ is the real-value Cartesian coordinates of its residue atoms (we only consider backbone atoms herein, *i.e.*, $[\text{N}, \text{C}_\alpha, \text{C}, \text{O}]$ with $N_{\text{atoms}} = 4$). Namely,

$$p_\theta(\text{prot}) = p_\theta(s_1, s_2, \dots, s_L,\ x_1, x_2, \dots, x_L) = p_\theta(\mathbf{s}, \mathbf{x})$$

Table 1: *Generative tasks w.r.t. structure & sequence.*

task	objective
folding	$p_\theta(\mathbf{x}\|\mathbf{s})$
inv-folding	$p_\theta(\mathbf{s}\|\mathbf{x})$
seq. gen.	$p_\theta(\mathbf{s})$
struct. gen.	$p_\theta(\mathbf{x})$
seq-struct co-gen.	$p_\theta(\mathbf{s}, \mathbf{x})$

As a result, most of protein tasks can be viewed as specifying their input conditioning and output between these two modalities (Tab. 1), including (1) sequence-conditioned structure prediction (folding, Jumper et al., 2021; Lin et al., 2022; Huguet et al., 2024), (2) structure-conditioned sequence generation (inverse folding or fixed-backbone design, Dauparas et al., 2022; Hsu et al., 2022; Zheng et al., 2023b), (3) sequence learning or generation (Rives et al., 2019; Nijkamp et al., 2022; Alamdari et al., 2023; Wang et al., 2024), (4) structure generation (Yim et al., 2023; Watson et al., 2023; Ingraham et al., 2023), and (5) sequence-structure co-generation (Jin et al., 2021; Shi et al., 2022;

Campbell et al., 2024). These further enable various conditional applications by allowing single or mixed-modal conditioning for partial generation, *e.g.*, motif-scaffolding and antibody design.

2.2 Diffusion Protein Language Model (DPLM)

Language models (LMs), typically parameterized by Transformers (Vaswani et al., 2017) have become the *de facto* choice dominating different domains with scalable and performing expressiveness (OpenAI, 2023). Among them, protein LMs have been serving as one of the AI foundation for protein sequence learning (Rives et al., 2019; Lin et al., 2022) and generation (Nijkamp et al., 2022; Alamdari et al., 2023).

Diffusion protein language model (DPLM, Wang et al., 2024), in particular, shows excelling performance in both generation and representation learning of protein sequences. DPLM is grounded in *absorbing* discrete diffusion framework (Austin et al., 2021; Zheng et al., 2023a), which is characterized by a forward and backward Markov process. Let $\texttt{Cat}(\mathbf{x}; \mathbf{p})$ be a categorical distribution on protein sequence $\mathbf{y}$ parameterized by a vector $\mathbf{p}$ on $(|\mathcal{V}|-1)$-dimensional probability simplex. The forward process of discrete diffusion defines a Markov process governed by the transition kernel $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \texttt{Cat}\big(\mathbf{x}^{(t)}; \beta_t \mathbf{x}^{(t-1)} + (1-\beta_t)\mathbf{q}_{\text{noise}}\big)$ that gradually perturb the data $\mathbf{x}^{(0)} \sim q(\mathbf{x}^{(0)})$ into a stationary distribution $\mathbf{x}^{(T)} \sim \mathbf{q}_{\text{noise}}$. For absorbing diffusion, $\mathbf{q}_{\text{noise}}$ is the point mass with all of the probability on the absorbing (mask) state. The learned *backward* process $p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ reversely denoises the $\mathbf{x}^{(T)}$ towards the data distribution $\mathbf{x}^{(0)}$, which is typically optimized by the variational bound of the log-likelihood (Ho et al., 2020):

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{x}^{(0)})}\big[\log p_\theta(\mathbf{x}^{(0)})\big] \geq \mathbb{E}_{q(\mathbf{x}^{(0:T)})}\left[\log \frac{p_\theta(\mathbf{x}^{(0:T)})}{q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\right] \\
&= \mathbb{E}_{q(\mathbf{x}^{(0)})}\Big[\log p_\theta(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}) + \textstyle\sum_{t=2}^{T} \underbrace{-\mathrm{KL}\big[q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) \| p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big]}_{\mathcal{J}_t}\Big] + \text{const.},
\end{aligned}$$

where $\mathcal{J}_t$ is the learning objective. The learning objective of discrete diffusion can be further simplified into reweighted cross-entropies (Zheng et al., 2023a), resembling masked language modeling at arbitrary noise levels:

$$\begin{aligned}
\mathcal{J}_t &= \mathbb{E}_{q(\mathbf{x}^{(0)})} - \mathrm{KL}\big[q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) \| p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big] \\
&= \mathbb{E}_{q(\mathbf{x}^{(0)})}\Big[\lambda^{(t)} \textstyle\sum_{1 \leq i \leq L} b_i(t) \cdot \log p_\theta(x_i^{(0)}|\mathbf{x}^{(t)})\Big],
\end{aligned} \tag{1}$$

where $\lambda^{(t)}$ is a weighting coefficient induced from the specific noising schedule and $b_i(t) = \mathbf{1}_{x_i^{(t)} \neq x_i^{(0)}}$. For inference, DPLM is able to generate amino acid sequences by the reverse iterative denoising process of discrete diffusion (Hoogeboom et al., 2021; Austin et al., 2021) from the following distribution,

$$p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \textstyle\sum_{\tilde{\mathbf{x}}^{(0)}} q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \tilde{\mathbf{x}}^{(0)}) p_\theta(\tilde{\mathbf{x}}^{(0)}|\mathbf{x}^{(t)}).$$

Specifically, at time t, it first generates $\tilde{\mathbf{x}}^{(0)}$ from $p_\theta(\cdot|\mathbf{x}^{(t)})$, then a less noisy $\mathbf{x}^{(t-1)}$ is sampled by $q(\cdot|\mathbf{x}^{(t)}, \mathbf{x}^{(0)} = \tilde{\mathbf{x}}^{(0)})$. Within absorbing diffusion, the generation process can be viewed as an iterative *mask-predict* approach. For sequence representation for predictive tasks, it can be obtained by simply letting DPLM take the sequence as input.

3 DPLM-2: A Multimodal Diffusion Protein Language Model

3.1 Overview

Fig. 1 illustrates DPLM-2's overall architecture. DPLM-2 is built on the state-of-the-art sequence-based generative protein LM, *i.e.*, DPLM (Wang et al., 2024), using a discrete diffusion probabilistic framework to concurrently model both protein sequences and their corresponding structures. To facilitate structure learning in language models, we introduce a token-based representation for protein structure via a tokenizer that converts $\mathbf{x} \in \mathbb{R}^{L \times N_{\text{backb}} \times 3}$, the 3D coordinates of the protein backbone into a discrete structure token sequence, denoted as $\mathbf{z} = (z_1, z_2, \ldots, z_L) \in \{0,1\}^{L \times |\mathcal{Z}|}$, where each token z_i represents a local structural element of the i-th residue. Given tokenized structure, DPLM-2 processes mulitmodal input by concatenating the structure token sequence $\mathbf{z}$ with the corresponding amino acid sequence $\mathbf{s}$ for the same protein. Notably, there exists a position-by-position correspondence between $\mathbf{z}$ and $\mathbf{s}$, where z_i and s_i refer to the two modalities of the i-th residue, respectively. To reinforce this correspondence, we assign identical position encodings to both z_i and s_i, thereby ensuring that structural and sequence information is aligned at the residue level.

To train DPLM-2, we leverage a high-quality dataset comprising 20K clustered experimental structures from the Protein Data Bank (PDB) (Berman et al., 2000) and 200K predicted structures from the AFDB SwissProt split (Varadi et al., 2022), with length < 512. During training, DPLM-2 is tasked with denoising the input sequence across a spectrum of noise levels, ranging from fully noisy to completely clean. The multimodal training objective of DPLM-2 is derived from Eq. (1) as,

$$\mathcal{J}_t = \mathbb{E}_{q(\mathbf{x}^{(0)},\mathbf{s}^{(0)}),\mathbf{z}^{(0)} \leftarrow \textit{tokenize}(\mathbf{x}^{(0)})} \Big[\lambda^{(t)} \textstyle\sum_{1 \le i \le L} b_i(t) \cdot \log p_\theta(z_i^{(0)}, s_i^{(0)} | \mathbf{z}^{(t)}, \mathbf{s}^{(t)}) \Big],$$

where $\log p_\theta(z_i, s_i|\cdot) = \log p_\theta(z_i|\cdot) + \log p_\theta(s_i|\cdot)$ by assuming conditional independence. By learning $p_\theta(\mathbf{z}^{(t-1)}, \mathbf{s}^{(t-1)}|\mathbf{z}^{(t)}, \mathbf{s}^{(t)})$, the model enables the simultaneous generation of highly correlated protein structures and sequences. This eliminates the need for a cascaded generation paradigm, allowing us to derive both the protein's structure and sequence in a single step.

To further enhance DPLM-2's ability to differentiate between structure and sequence, noising level for each modality is subjected to distinct scheduler, denoted as $t_\mathbf{z}$ and $t_\mathbf{s}$, respectively. This facilitates a more comprehensive understanding of the relationships between protein sequences and their corresponding structures. This design also allows us to explore arbitrary combinations of $(t_\mathbf{z}, t_\mathbf{s})$, thus providing flexible sampling options, including sampling from the marginals of each modality and conditionals between them for various applications (Fig. 1C). For conditional sampling (e.g., folding and inverse-folding), we set the noise scheduler of the conditioned modality to 0, which means no noise in the conditioned modality. For example, in the folding task, the t_s is always set to 0, while in the inverse-folding task the t_z is always set to 0. When sampling from the marginals of each modality, we set the noise scheduler of another modality to T, which is the maximum timestep and means 100% noise in another modality. For structure-sequence co-generation, we keep the t_z and t_s for the same to enhance the consistency between structure and sequence. Please refer to §A.4 for more details.

Furthermore, we also identify the exposure bias issue in discrete diffusion for sequence learning (Ranzato et al., 2016; Bengio et al., 2015), and mitigate this by proposing a self-mixup strategy inspired by scheduled sampling, which improves both generation quality and diversity (see §A.2).

3.2 Efficient Warm-up from Pre-trained Sequence-based DPLM

Protein sequences encode critical evolutionary information, reflecting co-evolutionary processes where residue pairs mutate together and often interact in 3D space, offering insights for predicting protein folding (Melnyk et al., 2022b). Lin et al. (2022) further showed that protein language models trained on large-scale evolutionary data implicitly capture this information, which can facilitate structure prediction. Motivated by the link between evolutionary knowledge and structural interactions, we propose to built DPLM-2 with an efficient warmup from pre-trained sequence-based DPLM, to make the most of established evolutionary information for protein structure modeling, Since our structure dataset is significantly smaller than UniRef50 sequence database (200K *vs.* 45M), enabling efficient fine-tuning of the pre-trained model. we want to keep the sequence knowledge intact and reduce the risk of catastrophic forgetting, we apply LoRA (Hu et al., 2021) to limit too much deviation to the original parameters. This approach not only lowers training costs compared to starting from scratch but also effectively transfers valuable evolutionary information.

3.3 Learning Structure Tokenization

The core difficulty of achieving a mulimodal protein LM lies in enabling the language model to learn structural information, which is challenging and remains elusive, Tokenizing continuous data modalities into discrete representations (Van Den Oord et al., 2017) has gained attraction across domains like image synthesis due to its ability to capture compact, meaningful information, enabling effective compression and efficient generation, especially with sequence-based models like Transformers. Recent efforts have applied this approach to protein structure coordinates (Van Kempen et al., 2024; Liu et al., 2023; Gao et al., 2024; Lu et al., 2024). This allows language models to better learn the composition of local structural elements. However, how to learn an effective structure tokenizer remains an active research question.

Structure tokenization under a typical VQ-VAE (Van Den Oord et al., 2017) framework can be summarized as follows:

$$\mathbf{x} \xrightarrow{\text{encoder}} \mathbf{e} \xrightarrow{\text{quantizer}} \mathbf{z} \xrightarrow{\text{decoder}} \tilde{\mathbf{x}},$$

where (1) a structure encoder encodes backbone 3D coordinates $\mathbf{x} \in \mathbb{R}^{L \times N_{\text{backb}} \times 3}$ into in-

A - reconstruction acc. of tokenizers

tokenizer	size	train	cameo 2022		
		lddt_ca	lddt_ca	tm-score	rmsd
VQ-VAE-1k	1024	0.76	0.71	0.80	6.14
LFQ-1k	1024	0.82	0.77	0.86	4.35
LFQ-2k	2048	0.84	0.79	0.88	3.62
LFQ-4k	4096	0.86	0.82	0.91	3.31
LFQ-8k	8192	0.92	0.86	0.93	2.58
LFQ-16k	16384	0.92	0.87	0.94	2.32

B - struct tokens *vs* secondary struct

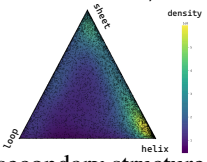


Figure 2: Reconstruction and secondary structure correspondence of structure tokenizers.

variant features $\mathbf{e} \in \mathbb{R}^{L \times d_{\text{quant}}}$, (2) a quantizer converts $\mathbf{e}$ into $\mathbf{z}$ of L discrete tokens where $z_i \in \{0, 1, \ldots, |\mathcal{Z}|\}$ given a finite-size codebook $\mathcal{Z}$; and (3) a structure decoder reconstructs 3D coordinates from the discrete tokens.

We utilize a GVP-based (Jing et al., 2020) structure encoder from pre-trained GVP-Transformer (Hsu et al., 2022) with its parameters frozen during training. The structure encoder transforms backbone structures into geometric features, which are projected onto a latent embedding using an MLP layer. The structure decoder follows the IPA-based modules from AlphaFold2 (Jumper et al., 2021), using 4 EvoFormer layers without MSA row attention, following ESMFold (Lin et al., 2022), to generate atomic positions from the structure tokens. We train structure tokenizer using the same structure data as our mulitmodal language model, containing both experimental and high-quality structures. The training objective of structure tokenizer includes reconstruction loss, codebook commitment loss, and entropy regularization loss to ensure effective codebook utilization. For the reconstruction loss, we adopt the FAPE loss, violation loss, and distogram loss from AlphaFold2 (Jumper et al., 2021), measuring the difference between predicted and native structures. To further enhance the training, we introduce a sequence prediction head on top of the structure decoder's final representation and minimize the cross-entropy against the native sequence.

In terms of quantizer, our preliminary experiment showed that conventional VQ-VAE pretty much struggles in training. To mitigate this, we instead adopts Lookup-Free Quantizer (LFQ) from the currently best visual tokenizer (Yu et al., 2023) to protein structure tokenization. Specifically, the latent space of LFQ is decomposed as the Cartesian product of single-dimensional binary variables, as $\mathbb{C} = \times_{k=1}^{\log_2 |\mathcal{Z}|} \mathcal{C}_k$, where $\mathcal{C}_k = \{-1, 1\}$. Given the encoded feature $\mathbf{e} = \text{encoder}(\mathbf{x}) \in \mathbb{R}^{L \times \log_2 |\mathcal{Z}|}$, each dimension (indexed by k) of the quantized representation $\text{quant}(e_i)$ is obtained from:

$$\text{quant}(e_i[k]) = \mathcal{C}_i = \text{sign}(e_i[k]) = -\mathbf{1}\{z_i[k] \leq 0\} + \mathbf{1}\{e_i[k] > 0\}.$$

As such, with LFQ, the token indices for $\mathbf{z} = \{z_1, z_2, ..., z_i, ..., z_L\}$ is given by:

$$z_i = \text{index}(e_i) = \textstyle\sum_{k=1}^{\log_2 |\mathcal{S}|} 2^{k-1} \mathbf{1}\{e_i[k] > 0\}, \ \forall z_i \in \mathbf{z}.$$

The LFQ-based structure tokenizer is trained on the same structure dataset as mentioned before, using a combination of reconstruction, commitment, and entropy regularization losses, similar to standard VQ-VAE.

Evaluation. As shown in Fig. 2A, LFQ significantly outperforms VQ-VAE regarding reconstruction accuracy while training of LFQ is much faster than VQ-VAE (2 *vs*. 15 days on 8 A100s). Increasing codebook size leads to improved reconstruction while a codebook size of 8192 achieves the best compression-reconstruction trade-off. Meanwhile in Fig. 2B, we observe a strong correlation between structure tokens and secondary structures. For instance, a lot of structure tokens concentrated at the alpha helix and beta sheet vertices, while some tokens lie between regions. This suggests that structure tokens the fine-grained structural elements in backbone local environment.

4 Experiments

In this section, we evaluate DPLM-2 on various generative and understanding scenarios, including unconditional protein generation (structure, sequence, and structure-sequence co-generation, §4.1), and a variety of conditional tasks, such as folding (§4.2), inverse folding (§4.3) and motif-scaffolding (§4.4), and a series of protein predictive tasks (§4.5).

4.1 Unconditional Protein Generation

The goal of unconditional protein generation is to produce both the 3D structure and amino acid sequence. Typically, this is done using a cascaded approach: either generating the structure first and then use another model to predict the sequence, or vice versa. Here, we focus on generating structure and sequence simultaneously. We evaluate DPLM-2 on both cascaded and simultaneous generation across three tasks: *unconditional structure generation*, *unconditional sequence generation*, and *structure-sequence co-generation*.

Following Multiflow (Campbell et al., 2024), we evaluate the generated proteins in terms of *quality*, *novelty* and *diversity*. **Designability** is measured through *self-consistency evaluation* and *foldability* (Yim et al., 2023; Watson et al., 2023; Wu et al., 2022a). Self-consistency evaluation is assessed by folding the generated sequence with ESMFold (Lin et al., 2022), then using `sc-TMscore` and `sc-RMSD` with the co-generated structure to evaluate similarity. Foldability is evaluated via ESMFold, with `pLDDT` > 70 considered plausible. **Novelty** is assessed by comparing generated structures to known ones in PDB using TMScore (`pdb-TM`), with lower values indicating greater novelty. **Diversity** is measured by calculating pairwise `TMscore` (`inner-TM`), where lower scores

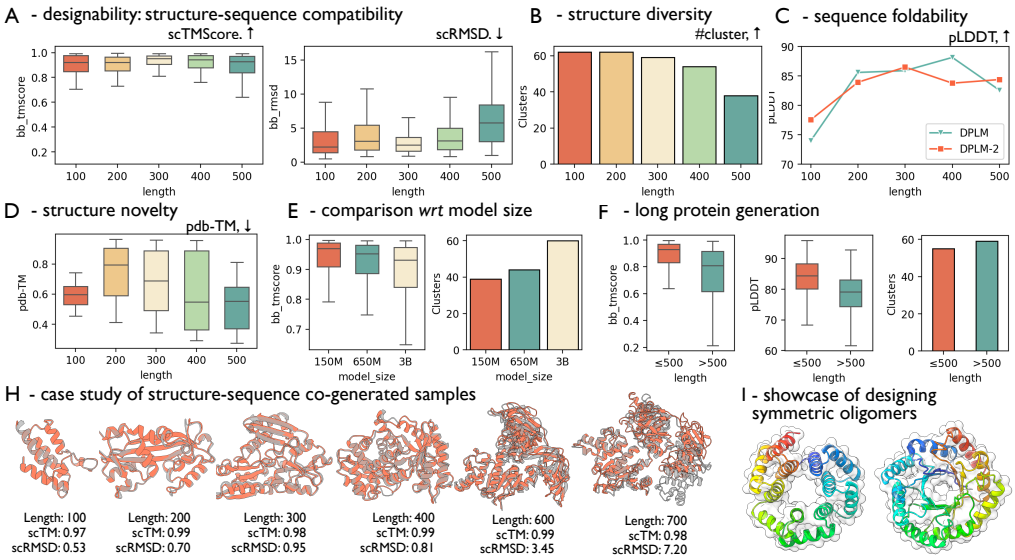


Figure 3: *Evaluation of unconditional structure-sequence co-generation.*

Table 2: Benchmarking comparison of unconditional protein generation.

	Designability			Novelty	Diversity	
	scTM (↑)	scRMSD (↓)	pLDDT (↑)	avg. pdb-TM (↓)	avg. inner-TM (↓)	MaxCluster (↑)
Structure-sequence co-generation.						
Native PDB protein	0.904 ± 0.129	4.623 ± 5.688	–	–	0.262 ± 0.025	0.776
ESM3 (seq → struct)	0.624 ± 0.232	24.180 ± 24.109	–	0.660 ± 0.000	0.220 ± 0.046	0.540
MultiFlow *w/* distillation (official ckpt)	0.930 ± 0.098	3.208 ± 4.741	79.447	0.704 ± 0.000	0.356 ± 0.032	0.500
*MultiFlow *w/o* distillation (official ckpt)	0.750 ± 0.163	9.306 ± 8.499	61.519	–	0.350 ± 0.038	0.490
*MultiFlow (retrained on our training data)	0.871 ± 0.934	6.580 ± 6.258	62.624	–	0.331 ± 0.052	0.440
DPLM-2 (struct → seq)	0.921 ± 0.098	4.969 ± 6.735	81.910	0.637 ± 0.195	0.308 ± 0.089	0.575
DPLM-2 (seq → struct)	0.907 ± 0.117	6.337 ± 9.403	82.246	0.653 ± 0.195	0.280 ± 0.038	0.651
DPLM-2 (co-generation)	0.925 ± 0.085	3.899 ± 3.723	82.686	0.640 ± 0.204	0.287 ± 0.030	0.545
Unconditional backbone generation. (sequence predicted by ProteinMPNN)						
Native PDB struct. (seq. from PMPNN)	0.969 ± 0.000	0.864 ± 0.000	–	–	0.262 ± 0.025	0.782
FrameDiff	0.818 ± 0.000	3.919 ± 0.000	–	0.668 ± 0.000	0.444 ± 0.064	0.252
FoldFlow	0.540 ± 0.000	7.965 ± 0.000	–	0.566 ± 0.000	0.286 ± 0.023	0.762
RFDiffusion	0.914 ± 0.000	1.969 ± 0.000	–	0.657 ± 0.000	0.352 ± 0.038	0.598
DPLM-2	0.945 ± 0.082	4.451 ± 5.261	–	0.637 ± 0.195	0.297 ± 0.049	0.575
Unconditional sequence generation. (structures predicted by ESMFold)						
EvoDiff	–	–	35.846	0.432 ± 0.106	0.265 ± 0.025	0.990
DPLM	–	–	83.252	0.541 ± 0.187	0.242 ± 0.041	0.735
DPLM-2	–	–	82.246	0.662 ± 0.199	0.280 ± 0.042	0.700

indicate more dissimilarity. The number of clusters identified by FoldSeek (van Kempen et al., 2023) also quantifies diversity, normalized by the total number of structures.

4.1.1 DPLM-2 Enables High-quality, Diverse and Novel Protein Sequence and Structure Generation

Tab. 2 and Fig. 3 present the results of DPLM-2 for unconditional protein generation. We highlight our key findings in the following aspects:

(1) DPLM-2 can generate diverse and highly-plausible protein with simultaneous structure-sequence co-generation. We sampled 100 proteins for each length in 100, 200, 300, 400, and 500. The co-generation can be performed in simultaneous generation (*co-generation*) and cascaded workflow: first generating the structure then the sequence conditioned on generated structure (*struct → seq*), and the reverse way (*seq → struct*), without the need of other folding or inverse folding models. Fig. 3A/B demonstrates that DPLM-2 can sample sequence and structures with high designability across various lengths, with most `sc-TM` values exceeding 0.9, with diverse structure clusters. Fig. 3D shows that the novelty of sampled proteins, measured by `pdb-TM`, generally increases with longer protein lengths. In addition, DPLM-2 can generate with both modalities simultaneously or a modality-by-modality. As shown in Tab. 2, the co-generation performance exhibit highest `scTM`, suggesting that co-modeling indeed benefits protein generation.

(2) DPLM-2 can attains competitive performance with strong baselines on co-generation, as well as backbone-only and sequence-only generation, respectively. As shown in Tab. 2, DPLM-2 achieves the strong `sc-TM` compared to strong baselines, approaching the quality of native structures

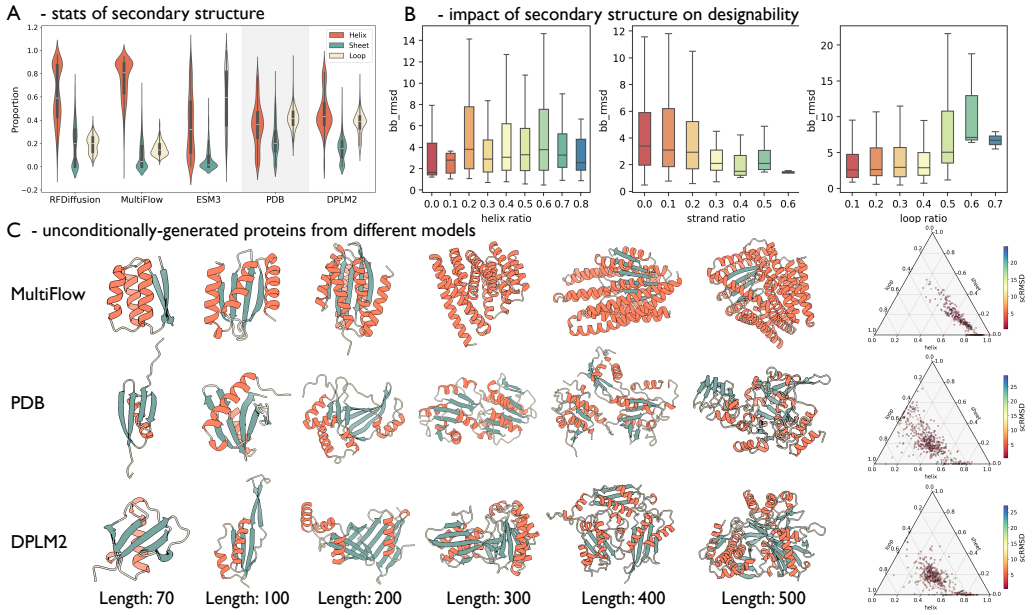


Figure 4: *Evaluation of secondary structure of generated proteins.*

from PDB. Compared to MultiFlow (Campbell et al., 2024), DPLM-2 achieves comparable co-generation quality. Notably, Multiflow's performance degrades greatly without data distillation from external inverse folding models, while we also provide the result of Multiflow retrained using our training data for reference. We also notice that ESM3 (Hayes et al., 2024), which runs in a sequence-then-structure order, fails short of unconditional generation. Moreover, DPLM-2 can also only produce single modality if needed, where it matches the best competitive models in these settings respectively. These results demonstrate DPLM-2's effectiveness as a mulitmodal generative model.

(3) DPLM-2 generates longer proteins beyond training data. We sample proteins at lengths of $[600, 700, 800, 900, 1000]$. As shown in Fig. 3F, notably, for proteins exceeding the maximum training length of 512, the `pLDDT` scores of sequences sampled by DPLM-2 are close to those of DPLM. This suggests that DPLM-2 largely retains its original sequence generation capability without suffering from catastrophic forgetting, demonstrating its capability of length extrapolation.

(4) Case study. Fig. 3H shows some generated samples of DPLM-2 up to 700 residues, while in Fig. 3I we showcase that we can manipulate DPLM-2 to design symmetric oligomers by forcing to duplicate the predicted tokens with repetitive structure and sequence patterns.

(5) Abaltion study on the training strategy. We investigate the effects of warmup from the sequence-based pre-trained DPLM and data augmentation with high-quality AlphaFold-predicted structures on DPLM-2. The sequence pre-training significantly improve both designability and diversity, while data augmentation can further enhance the designability, especially for long proteins. For more details of ablation study, please refer to §A.1.

4.1.2 DPLM-2 Generates Proteins That Resembles Natural Proteins

To further analyze the properties of different model, we examine their secondary structure distribution against natural proteins from PDB.

Proteins sampled by DPLM-2 have secondary structures most similar to natural proteins. As seen in Fig. 4A, structure-based models like RFDiffusion and MultiFlow generate proteins with more helices and fewer sheets and loops than natural proteins in PDB. Protein language models like ESM3 and DPLM-2 show no strong bias towards alpha helices, but ESM3 tends to generate more loops. Among the methods, DPLM-2 produces the most natural-like secondary structure proportions, closely matching PDB proteins. In Fig. 4C, proteins generated by MultiFlow contain many helices and become more globular as length increases, exhibiting idealized secondary structures. In contrast, proteins generated from DPLM-2 resembles natural ones have more balanced structures, with fewer helices and more beta sheets and loops. On the other hands, simplex plots in Fig. 4C shows that while MultiFlow's proteins are clustered in helix-rich regions, DPLM-2's proteins span a wider area similar to natural proteins, while it rarely samples proteins composed mostly of sheets and loops, which do occur in nature. Additionally, Fig. 4B shows that the loop ratio has a significant impact on designability, where a higher proportion of loops will increase `scRMSD`, as loops are highly

flexible. Thus, proteins with long loops, which DPLM-2 often generates, tend to have relatively high scRMSD, aligning with the results in Tab. 2.

4.2 Folding (Sequence-conditioned Structure Prediction)

The goal of folding is to predict the 3D structure for the given amino acid sequence (Jumper et al., 2021). As a mulitmodal generative model, DPLM-2 spontaneously enables protein structure prediction task (see Fig. 1C-3) given sequence as conditioning. We assess DPLM-2 on CAMEO 2022 and a PDB data split used by Multiflow (Campbell et al., 2024). We utilize RMSD and TMscore between predicted and ground truth structure for evaluation, while DPLM-2 adopts argmax decoding for 100 sampling iterations.

Tab. 3 indicates that DPLM-2 can perform sufficiently good folding in a zero-shot manner. Performance can be improved after further supervised fine-tuning (SFT) using folding objective ($\max_\theta \log p_\theta(\mathbf{z}|\mathbf{s})$). Overall, DPLM-2 can outperform or on par with the strong baselines, while achieving close performance with ESMFold. Plus, we observe that DPLM-2 with larger model scales can attain better results than smaller ones. We suggest that DPLM-2 benefits from the evolutionary information inherited from DPLM pre-trained on the vast number of protein sequences, which can be transferred and leveraged into structure modeling.

Table 3: *Structure prediction performance comparison between* DPLM-2 *and different baseline approaches on CAMEO 2022 datasets.* †: PVQD results are quoted from Liu et al. (2023).

Models	CAMEO 2022		PDB date split	
	RMSD	TMscore	RMSD	TMscore
ESMFold	3.99/2.03	0.85/0.93	2.84/1.19	0.93/0.97
†PVQD	4.08/1.95	0.81/0.88	–	–
MultiFlow	17.84/17.96	0.50/0.46	15.64/16.08	0.53/0.49
ESM3	6.33/2.98	0.85/0.92	4.94/2.28	0.87/0.93
DPLM-2 (150M)	9.22/7.64	0.75/0.81	8.35/5.60	0.76/0.82
w/ folding SFT	7.66/4.37	0.80/0.86	6.00/3.41	0.83/0.88
DPLM-2 (650M)	7.37/4.89	0.79/0.86	5.67/3.33	0.83/0.88
w/ folding SFT	6.21/3.78	0.84/0.89	3.40/1.78	0.89/0.94
DPLM-2 (3B)	6.34/3.65	0.83/0.89	4.54/2.54	0.86/0.92
w/ folding SFT	5.71/3.23	0.85/0.90	3.15/1.69	0.90/0.95

4.3 Inverse Folding (Structure-conditioned Sequence Generation)

The goal of inverse folding is to find an amino acid sequence that can fold to a given backbone structure. For evaluation, we employ amino acid recovery (AAR) for sequence evaluation, and we also assess the structure by self-consistency TM-score (scTM) between the native structure and the ESMFold-predicted structure of the generated sequence.

DPLM-2 can generate reasonable sequences that fold into the given structures. Tab. 4 presents that DPLM-2 can outperform or be on par with other co-generation models (MultiFlow, ESM3). As the model size increases, the performance in terms of sequence recovery (AAR) and structural consistency (scTM) improves, revealing the same scaling law observed in the folding task. We suggest that multimodal training effectively aligns the structure and sequence into the same space, such that DPLM-2 can yield the corresponding sequence without additional training.

Table 4: *Comparison on inverse folding task.*

Models	CAMEO 2022		PDB date split	
	AAR	scTM	AAR	scTM
MultiFlow	32.28/33.58	0.87/0.94	37.74/37.59	0.94/0.96
ESM3	47.06/46.24	0.90/0.95	49.50/49.42	0.94/0.97
DPLM-2 (150M)	45.22/46.12	0.87/0.93	48.83/47.96	0.89/0.95
DPLM-2 (650M)	49.01/50.10	0.88/0.93	54.80/53/07	0.91/0.96
DPLM-2 (3B)	52.36/53.72	0.89/0.95	61.67/57.91	0.92/0.96

4.4 Scaffolding with Mixed-modal Motif Conditioning

The objective of motif-scaffolding is to generate a suitable scaffold to preserve the structure of the given motif and maintain its original function. We follow the experimental setting of Yim et al. (2024), with 24 motif-scaffolding problems and we sample 100 scaffolds for each motif, where we (1) first determine the length of scaffold, and then (2) keep the motif segment unchanged and sample the scaffold part conditioned on the motif. The scaffold length is sampled from a range provided by Yim et al. (2024), and when there are multiple motifs, the order of motif segments is consistent with Yim et al. (2024). We provide the 3D structure and sequence of motif as input of DPLM-2. As a multimodal model, we evaluate DPLM-2 using sequence-based, structure-based, and co-generation approaches. A scaffold is considered successful if it satisfies both criteria (1) overall designablity, which is successful when pLDDT > 70 (for sequence-based models) or scTM > 0.8, and (2) motif-preseving, which is deemed successful when the predicted motif structure matches the native one with motif-RMSD <1Å.

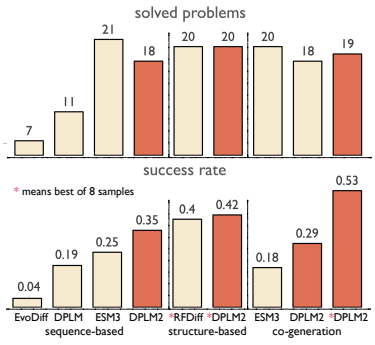


Figure 5: *Evaluation of motif-scaffolding w.r.t. success rate and num. of solved problems.*

Fig. 5 reveals that DPLM-2 is capable of generate reasonable scaffolds for the given functional motifs. In sequence-based, structure-based and co-generation evaluation, DPLM-2 can outperform or be on par with the corresponding approaches in most cases, solving more motif problem and achieving higher average success rate. We compared to sequence-based method, DPLM-2 shows better performance since it allows structural input of motif, which is important for preserving motif's structure hence the functions. Remarkably, DPLM-2 attains comparable performance with RFDiffusion when only generating scaffold structure, while achieve better performance when simultaneously designing scaffold sequence and structure, outperforming ESM3. Despite not experimentally verified, these results suggest that with DPLM-2, mulitmodal conditioning and generation could lead to more successful conditional protein design.

4.5 Evaluation of Protein Representation Learning

Directly access to structure information is supposed to benefit downstream protein predictive tasks. To inspect this, we evaluate DPLM-2 on a variety of protein predictive tasks utilizing the dataset provided by SaProt (Su et al., 2023), where we provide tokenized protein structure tokens along with the protein sequences to DPLM-2.

Table 5: *Performance on various protein predictive downstream tasks.* †: benchmarked results are quoted from Su et al. (2023).

Models	Thermostability	HumanPPI	Metal Ion Binding	EC	GO			DeepLoc	
					MF	BP	CC	Subcellular	Binary
	Spearman's ρ	Acc (%)	Acc (%)	Fmax	Fmax	Fmax	Fmax	Acc (%)	Acc (%)
†SaProt (650M)	**0.724**	**86.41**	**75.75**	**0.884**	*0.678*	0.356	**0.414**	**85.57**	*93.55*
†MIF-ST (Yang et al., 2022b)	0.694	75.54	75.08	0.803	0.627	0.239	0.248	78.96	91.76
†GearNet (Zhang et al., 2023)	0.571	73.86	71.26	0.871	0.650	0.354	0.404	69.45	89.18
ESM2 (650M)	0.691	*84.78*	71.88	0.866	0.676	0.344	0.402	83.68	92.28
DPLM (650M)	0.695	**86.41**	*75.15*	0.875	**0.680**	*0.357*	0.409	*84.56*	93.09
DPLM-2 (650M)	*0.714*	84.44	74.28	*0.878*	**0.680**	**0.359**	*0.411*	82.98	**93.64**

DPLM-2 can perform multimodal representation learning by leveraging both structure and sequence information. Tab. 5 presents that DPLM-2 shows further improvement compared to sequence-only methods (ESM2, DPLM) on some tasks, indicating that DPLM-2 can leverage protein structures to generate better representations containing multimodal information for downstream tasks. However, we find that DPLM-2 falls behind the state-of-the-art structure-aware protein LM, *i.e.*, SaProt, in most tasks and even lags behind DPLM in certain tasks. We hypothesize this is because the strutcure training data of DPLM-2, consisting of PDB and SwissProt, is smaller and differs from UniRef50, which DPLM is pretrained on, potentially causing catastrophic forgetting and suboptimal representation. To test this, we conducted an experiment on the DeepLoc subcellular task, where DPLM-2 underperforms compared to DPLM. As shown in Tab. 6, without large-scale sequence pretraining, DPLM-2 outperforms DPLM significantly, suggesting that: (1) Incorporating structure information enhances performance over sequence-only models. (2) Smaller datasets can lead to catastrophic forgetting, diminishing the benefits of large-scale pretraining. As result, to further improve the predictive performance, one deserving direction is to exploit larger-scale predicted structures in our future work.

Table 6: *Performance without large-scale sequence pre-training.*

Models	DeepLoc Subcellular
	Acc (%)
DPLM (650M)	63.49
DPLM-2 (650M)	**66.77**

5 Discussion

In this paper, we introduce DPLM-2, a diffusion protein language model that understands, generates and reasons over structure and sequence, aiming to severe as a mulimodal foundation for protein. Despite promising performance spanning protein co-generation, folding, inverse folding and conditional motif-scaffolding with mulimodal input and output, there remains several limitations deserving to be addressed. (1) Structure data: Our findings indicate that while structure awareness may help with predictive tasks, the limited structure data constrains DPLM-2's ability to learn robust representations. It is also important to account for longer protein chains and multimers in future studies. (2) Trade-off of discrete latent representation: Tokenizing structure into discrete symbols facilitates multimodal protein language models and co-generation but comes at the cost of losing fine-grained structural details and control, such as precise atomic positions and inter-atomic distances. Future work should aim to integrate the strengths of data-space structure-based generative models with sequence-based language models to maximize the best of both worlds.

REFERENCES

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

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 1171–1179, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

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

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

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

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

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

Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z Li. Foldtoken: Learning protein language via vector quantization and beyond. *arXiv preprint arXiv:2403.09673*, 2024.

Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pp. 2024–07, 2024.

Liang He, Shizhuo Zhang, Lijun Wu, Huanhuan Xia, Fusong Ju, He Zhang, Siyuan Liu, Yingce Xia, Jianwei Zhu, Pan Deng, et al. Pre-training co-evolutionary protein representation via a pairwise masked language model. *arXiv preprint arXiv:2110.15527*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al. Sequence-augmented se (3)-flow matching for conditional protein backbone generation. *arXiv preprint arXiv:2405.20313*, 2024.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, pp. 1–9, 2023.

Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi S Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. In *International Conference on Learning Representations*, 2021.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2020.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation. *arXiv preprint arXiv:2208.06073*, 2022.

Jin Sub Lee, Jisun Kim, and Philip M Kim. Proteinsgm: Score-based generative modeling for de novo protein design. *bioRxiv*, pp. 2022–07, 2022.

Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *arXiv preprint arXiv:2301.12485*, 2023.

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

Haiyan Liu, Yufeng Liu, and Linghui Chen. Diffusion in a quantized vector space generates non-idealized protein structures and predicts conformational distributions. *bioRxiv*, pp. 2023–11, 2023.

Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *BioRxiv*, pp. 2020–09, 2020.

Amy X Lu, Wilson Yan, Kevin K Yang, Vladimir Gligorijevic, Kyunghyun Cho, Pieter Abbeel, Richard Bonneau, and Nathan Frey. Tokenized and continuous embedding compressions of protein sequence and structure. *bioRxiv*, pp. 2024–08, 2024.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Deep neural language modeling enables functional protein generation across families. *bioRxiv*, pp. 2021–07, 2021.

Matthew McDermott, Brendan Yap, Harry Hsu, Di Jin, and Peter Szolovits. Adversarial contrastive pre-training for protein sequences. *arXiv preprint arXiv:2102.00466*, 2021.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, pp. 29287–29303, 2021.

Igor Melnyk, Vijil Chenthamarakshan, Pin-Yu Chen, Payel Das, Amit Dhurandhar, Inkit Padhi, and Devleena Das. Reprogramming large pretrained language models for antibody sequence infilling. *arXiv preprint arXiv:2210.07144*, 2022a.

Igor Melnyk, Aurelie Lozano, Payel Das, and Vijil Chenthamarakshan. Alphafold distillation for improved inverse protein folding. *arXiv preprint arXiv:2210.03488*, 2022b.

Seonwoo Min, Seunghyun Park, Siwon Kim, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Pre-training of deep bidirectional protein sequence representations with structural information. *IEEE Access*, 9:123912–123926, 2021.

Ananthan Nambiar, Maeve Heflin, Simon Liu, Sergei Maslov, Mark Hopkins, and Anna Ritz. Transforming the language of life: transformer neural networks for protein prediction tasks. In *Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics*, pp. 1–8, 2020.

Erik Nijkamp, Jeffrey Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *arXiv preprint arXiv:2206.13517*, 2022.

Esmaeil Nourani, Ehsaneddin Asgari, Alice C McHardy, and Mohammad RK Mofrad. Tripletprot: deep representation learning of proteins based on siamese networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3744–3753, 2021.

OpenAI. Gpt-4 technical report, 2023.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06732.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v4.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Nils Strodthoff, Patrick Wagner, Markus Wenzel, and Wojciech Samek. Udsmprot: universal deep sequence models for protein classification. *Bioinformatics*, 36(8):2401–2409, 2020.

Pascal Sturmfels, Jesse Vig, Ali Madani, and Nazneen Fatema Rajani. Profile prediction: An alignment-based pre-training task for protein sequence models. *arXiv preprint arXiv:2012.00195*, 2020.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, pp. 2023–10, 2023.

Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Serbulent Unsal, Heval Atas, Muammer Albayrak, Kemal Turhan, Aybar C Acar, and Tunca Doğan. Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4 (3):227–245, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, pp. 1–4, 2023.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, volume 30, pp. 5998–6008, 2017.

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

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024.

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

Kevin E Wu, Kevin K Yang, Rianne van den Berg, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *arXiv preprint arXiv:2209.15611*, 2022a.

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

Yijia Xiao, Jiezhong Qiu, Ziang Li, Chang-Yu Hsieh, and Jie Tang. Modeling protein using large-scale pretrain language model. *arXiv preprint arXiv:2108.07435*, 2021.

Kevin K Yang, Alex X Lu, and Nicolo Fusi. Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv*, pp. 2022–05, 2022a.

Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *bioRxiv*, pp. 2022–05, 2022b.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Jason Yim, Andrew Campbell, Emile Mathieu, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Frank Noé, et al. Improved motif-scaffolding with se (3) flow matching. *arXiv preprint arXiv:2401.04082*, 2024.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*, 2023.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*, 2023a.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei YE, and Quanquan Gu. Structure-informed language models are protein designers. In *International Conference on Machine Learning*, 2023b.

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference optimization. *Advances in neural information processing systems*, 2024.

A DPLM-2 Training

A.1 Ablation Study

In DPLM-2 training, we start with a warmup from the sequence-based pre-trained DPLM to exploit established evolutionary information and augment the data with high-quality AlphaFold-predicted structures from SwissProt (around 200K) and clustered PDB structures. This section evaluates the effects of sequence pre-training and data augmentation on unconditional protein generation.

Table 7: Ablation study on the sequence pre-training and training data augmentation.

sequence pre-training	synthetic structures	length 100		length 200		length 300		length 400		length 500	
		scTM	clusters	scTM	clusters	scTM	clusters	scTM	clusters	scTM	clusters
✗	✗	0.9241	20	0.8674	34	0.7667	33	0.5016	25	0.4511	25
✓	✗	**0.9610**	26	0.9349	**47**	0.9169	38	0.8643	35	0.7673	**52**
✗	✓	0.8988	27	0.9182	15	**0.9343**	13	0.8518	21	0.8288	31
✓	✓	0.9348	**35**	**0.9428**	40	0.9232	**48**	**0.9260**	**40**	**0.9012**	32

We investigate the effect of sequence pre-training by randomly initializing DPLM-2 instead of using DPLM parameters, while for effect of synthetis structures we leverage PDB structures only for training. We conduct experiments on 150M DPLM-2, for each DPLM-2 variant we sample 100 examples for each length in 100, 200, 300, 400 and 500. We compute `scTM` and the number of difference clusters in each length. Tab. 7 demonstrates that *sequence pre-training and data augmentation can significantly improve the designability and diversity*, especially in generating long proteins (length > 300). We hypothesize that the limited number of long proteins in PDB leads to insufficient training. In contrast, sequence pretraining, which includes evolutionary data, is essential and can be transferred to improve protein structure modeling and generation quality. Additionally, this evolutionary information boosts sampling diversity. While increasing the amount of training data improves designability, it is less effective in enhancing diversity compared to sequence pretraining. By combining both strategies, we achieve the best overall performance, which forms the core of our training strategy.

A.2 Self-mixup training strategy

We find that discrete diffusion training will face the *exposure bias* problem (Ranzato et al., 2016; Bengio et al., 2015), which means mismatch between training and inference. The model is trained to denoise given the ground-truth context during training. However, during inference, the model needs to denoise based on the predicted tokens, which may not be correct and inconsistent with the always-accurate context during training. This may lead to error accumulation and negatively impact the generation performance.

To address this issue, we propose a *self-mixup* training paradigm for discrete diffusion model, enhancing the consistency between training and inference. During training, we perform an additional forward pass, allowing the model to first make predictions and then denoise based on those predictions.

Tab. 8 shows that the *self-mixup* training strategy effectively enhances the diversity of samples. We attribute this to the model producing more accurate logits during inference, leading to more diverse reasonable sampling paths instead of converging on the sampling paths with the highest probability, which results in more diverse proteins.

Table 8: Ablation study on the *self-mixup* training strategy.

Mixup strategy	length 100		length 200		length 300		length 400		length 500	
	scTM	clusters	scTM	clusters	scTM	clusters	scTM	clusters	scTM	clusters
✗	**0.9237**	44	**0.9180**	53	0.9147	48	0.9059	42	**0.8896**	33
✓	0.8812	**62**	0.8820	**62**	**0.9172**	**59**	**0.9099**	**54**	0.8845	**38**

A.3 Implementation Details

DPLM-2 takes the discrete structure token sequence and amino acid token sequence as input. As demonstrated in Fig. 1, we concatenate the two sequences into one sequence of double length. DPLM-2 employs an efficient warm-up strategy by initializing with pre-trained sequence-based DPLM (§3.2) to leverage the evolutionary information learned by DPLM for protein structure modeling. Considering that the vocabulary of DPLM only consists of amino acids, we expand the vocabulary of DPLM-2 with discrete structure tokens. We initialize the embeddings of structure tokens with the mean and standard variation of the learned amino acid embeddings. We hypothesis this will keep the initial embedding distribution remains consistent with the pre-trained DPLM, resulting in more stable training in early stage and preventing excessive gradients that could lead to training crashes.

A.4 Distinct Noise Scheduler of Training

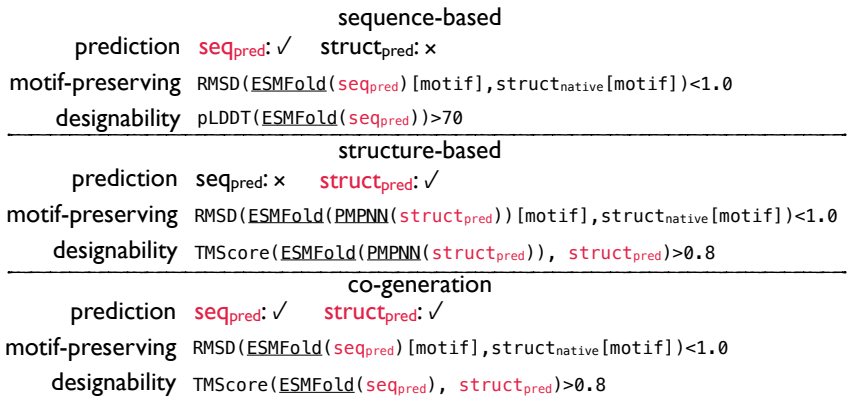
We introduce a distinct scheduler to control the noise level of structure and sequence flexibly during training (§3.1). Different combinations of structures and sequence schedulers (denoted as t_z and t_s, respectively) imply training for different applications. Specifically, we mainly focus on (1) sequence-conditioned structure generation (e.g., folding), (2) structure-conditioned sequence generation (e.g., inverse-folding), (3) sequence generation, (4) structure generation, (5) structure-sequence co-generation, as shown in Tab. 1. For conditional generation tasks (e.g., folding and inverse-folding), we set the noise scheduler of the conditioned modality to 0, e.g., no noise in the conditioned modality. Specifically, in the folding task, the t_s is always set to 0, while in the inverse-folding tasks the t_z is always set to 0. In the structure-sequence co-generation task, we keep the t_z and t_s for the same, enhancing the structure-sequence consistency in co-generation. The structure or sequence generation tasks do not depend on another modality, so we set the noise scheduler of another modality to T, e.g., 100% noise in another modality. For example, in structure generation task, the t_s is always set to T.

During training, we train the above 5 tasks simultaneously. We divide the training data in a batch into 5 parts according to a preset proportion, and each part is used for a specific task training. In our experiment, the proportion for each task is the same, which is 20%. After training, we can further enhance a specific generation task by supervised finetuning (SFT). This involves continuing training for the specific task with a proportion of 100%, while the proportion for other tasks is set to 0%. For example, in Tab. 3, the folding supervised finetuning is performed by continue training based on a pre-trained DPLM-2 with 100% proportion of training data.

A.5 Dataset

The training set of DPLM-2 is composed by experimental data, *i.e.*, PDB (Berman et al., 2000), and high quality synthetic data, *i.e.*, SwissProt (Varadi et al., 2022). We filter the SwissProt data by pLDDT > 85. After filtering, the overall training set contains approximately 200,000 proteins. We limit the maximum length of the training set to 512. For proteins longer than 512, we randomly crop it to 512. We crop the low pLDDT (pLDDT < 50) segments located at the both ends of proteins in the SwissProt dataset. These segments are typically non-structural and may negatively impact the training results. Moreover, we find that the length distribution of the training set is not balanced, where the number of proteins with length less than 100 is relatively small, leading to a suboptimal diversity among the short proteins. Therefore, during training, we randomly crop long proteins to short proteins with a probability of 50% for each batch to improve the diversity.

Figure 6: Sequence-based, structure-based and co-generation evaluation pipeline of motif-scaffolding.



A.6 Hyperparameter

We train all models using AdamW optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We use a weight decay of 0.01 and gradient clipping of 0.5. We employ 2K warmup steps until reaching the maximum learning rate, and utilize a linear decay scheduler to decay LR to 10% of the maximum learning rate by the end of training. The maximum learning rate is 1e-4, and the overall training step is 100,000. We utilize the pretrained DPLM as the parameter initialization, and the diffusion timestep is set to 500. We train 150M DPLM-2 with 8 A100 GPUs for 3 days, while 650M with 16 A100 GPUs for 3 days and 3B with 16 A100 GPUs for a week.

B Motif Scaffolding

B.1 Evaluation Pipeline

We evaluate DPLM-2 in sequence-based, structure-based and co-generation ways. The overall illustration is shown in Fig. 6.

We focus on the two aspects: overall quality and motif part consistency. The assessment of overall quality varies across different approaches. Specifically, (1) For sequence-based method, we only take the generated sequence and utilize ESMFold to obtain the predicted structure, and the `pLDDT` score provided by ESMFold is used to assess overall quality. (2) For structure-based method, we only take the generated structure, and then leverage ProteinMPNN to predict the sequence, followed by ESMFold to predict the structure, where overall quality is assessed by `scTM`. (3) For co-generation method, we take both the generated structure and sequence, and predict structure given generated sequence with ESMFold, where `scTM` is calculated between generated structure and ESMFold predicted structure to evaluate overall quality. Considering that the ground truth motif structure is given, we only utilize the ESMFold predicted structure to calculate `motif-RMSD`.

B.2 Result of Each Problem

Tab. 9 presents the result of each motif-scaffolding problem. DPLM-2 achieves the best average success rate in each evaluation. Compared with ESM3, DPLM-2 shows better results in 12 problems in co-generation evaluation and 10 problems in sequence-based evaluation. Meanwhile, DPLM-2 outperforms RFDiffusion in 14 problems in structure-based evaluation. This demonstrates that DPLM-2 can achieve strong performance under various evaluation methods.

We also find that taking the best result from 8 samples can bring significant improvement compared to 1 sample, especially in terms of success rate. In the co-generation evaluation, DPLM2 with sampling 8 times improves the success rate of most of the problems by a large margin. We hypothesize that sampling eight times largely alleviates errors caused by randomness in the sampling process, thereby producing a more suitable scaffold for the given motif.

C Related Work

C.1 Protein Language Models

There is growing interest in developing protein LMs at the scale of evolution, such as the series of ESM (Rives et al., 2019; Lin et al., 2022), TAPE (Rao et al., 2019), ProtTrans (Elnaggar et al., 2021), PRoBERTa (Nambiar et al., 2020), PMLM (He et al., 2021), ProteinLM (Xiao et al., 2021),

Table 9: Motif-scaffolding results of each problem. * means best result from 8 samples.

	sequence-based				structure-based		co-generation		
	EvoDiff	DPLM	ESM3	DPLM2	*RFDiffusion	*DPLM2	ESM3	DPLM2	*DPLM2
1BCF	0.00	0.00	**0.89**	0.01	**1.00**	0.07	**0.23**	0.01	0.05
1PRW	0.61	0.83	**0.96**	0.86	0.08	**0.96**	0.54	0.84	**0.95**
1QJG	0.00	0.00	0.02	**0.03**	0.00	0.00	0.03	0.02	**0.05**
1YCR	0.02	0.38	0.41	**0.77**	0.74	**0.93**	0.18	0.53	**0.98**
2KL8	0.04	0.08	0.11	**0.47**	0.88	**0.94**	0.11	0.57	**1.00**
3IXT	0.06	0.17	0.18	**0.67**	0.25	**0.77**	0.02	0.41	**0.73**
4JHW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4ZYP	0.00	0.00	0.03	**0.16**	0.40	**0.51**	0.08	0.10	**0.64**
5IUS	0.00	0.00	0.00	0.00	**0.02**	0.00	0.00	0.00	0.00
5TPN	0.00	0.00	**0.03**	0.00	**0.61**	0.06	**0.01**	0.00	0.00
5TRV_long	0.00	0.00	**0.19**	0.00	**0.37**	0.08	**0.19**	0.00	0.07
5TRV_med	0.00	0.00	**0.16**	0.03	**0.24**	0.07	0.16	0.02	**0.19**
5TRV_short	0.00	0.00	0.01	**0.07**	0.04	**0.10**	0.01	0.03	**0.11**
5WN9	0.00	0.00	**0.02**	0.00	0.00	**0.20**	0.00	0.00	0.00
5YUI	0.00	0.00	0.00	0.00	**0.02**	0.00	0.00	0.00	0.00
6E6R_long	0.01	0.65	0.07	**0.91**	0.86	**0.92**	0.04	0.78	**1.00**
6E6R_med	0.03	**0.94**	0.24	0.93	**0.89**	0.88	0.14	0.77	**0.97**
6E6R_short	0.07	**0.87**	0.09	0.86	0.39	**0.78**	0.06	0.64	**0.99**
6EXZ_long	0.00	0.01	0.32	**0.61**	**0.76**	0.63	0.13	0.44	**0.95**
6EXZ_med	0.00	0.00	0.31	**0.66**	0.49	**0.63**	0.31	0.55	**0.96**
6EXZ_short	0.00	0.00	0.31	**0.66**	0.39	**0.41**	0.28	0.58	**0.87**
7MRX_long	0.00	0.02	**0.36**	0.23	0.09	**0.32**	0.37	0.20	**0.73**
7MRX_med	0.00	0.31	**0.65**	0.28	0.11	**0.31**	0.59	0.22	**0.70**
7MRX_short	0.00	0.34	**0.68**	0.26	0.02	**0.41**	0.74	0.24	**0.88**
pass rate	7/24	11/24	**21/24**	18/24	20/24	20/24	**20/24**	18/24	19/24
avg. success rate	0.04	0.19	0.25	**0.35**	0.40	**0.42**	0.18	0.29	**0.53**

PLUS (Min et al., 2021), Adversarial Masked LMs (McDermott et al., 2021), ProteinBERT (Brandes et al., 2022), CARP (Yang et al., 2022a) in masked language modeling (MLM) paradigm, ProtGPT2 (Ferruz et al., 2022) in causal language modeling paradigm, and several others (Melnyk et al., 2022a; Madani et al., 2021; Unsal et al., 2022; Nourani et al., 2021; Lu et al., 2020; Sturmfels et al., 2020; Strodthoff et al., 2020). These protein language models exhibit remarkable generalization ability on various downstream tasks and be able to capture evolutionary information about secondary and tertiary structures from sequences alone. Meanwhile, recent study shows these models' potency in revealing protein structures (Lin et al., 2022), predicting the effect of sequence variation on function (Meier et al., 2021), antibody infilling (Melnyk et al., 2022a) and many other general purposes (Rives et al., 2019). Simultaneously, Verkuil et al. (2022) demonstrate that the large scale protein LMs can generate *de novo* proteins by generalizing beyond natural proteins, both theoretically and experimentally validating their hypothesis in exhaustive detail, in which *p*LMs demonstrate competency in designing protein structure despite being exclusively trained on sequences.

C.2 Protein Structure Generative Models

Diffusion models have become popular tools in structural biology for protein generation, and their utility has been demonstrated across a range of generative tasks in recent years. Trippe et al. (2022), along with others, have introduced several diffusion model variants, each with its unique approach. For instance, while some models focus on generating the protein backbone by diffusing over protein coordinates, others, such as those proposed by Wu et al. (2022b), target inter-residue angles. Lin & AlQuraishi (2023) and Yim et al. (2023) have developed models that handle both the position and orientation of residue frames. RFDiffusion (Watson et al., 2023) is a model that assists in designing protein structures for specific functions, such as enzymes. It is versatile in protein design and has been used to create therapeutic proteins, with some designs being confirmed in the laboratory. ProteinSGM (Lee et al., 2022) is a model that uses 2D matrices, which represent the distances and angles between protein parts, to create 3D protein structures for novel protein designs. FoldingDiff (Wu et al., 2022a) is a model that generates protein sequences expected to fold into a specific structure. These sequences are verified with prediction tools, although they have not been experimentally confirmed yet. Chroma (Ingraham et al., 2023) is a model designed for creating large proteins and protein complexes, considering various constraints like distances and symmetry. It transforms a collapsed polymer into protein backbone and sequence more quickly than older methods, thereby allowing for the efficient generation of large structures. Multiflow (Campbell et al., 2024) develop mulitmodal flow matching for protein structure-sequence co-generation (Jin et al., 2021; Shi et al., 2022).