MSAGPT: Neural Prompting Protein Structure Prediction via MSA Generative Pre-Training

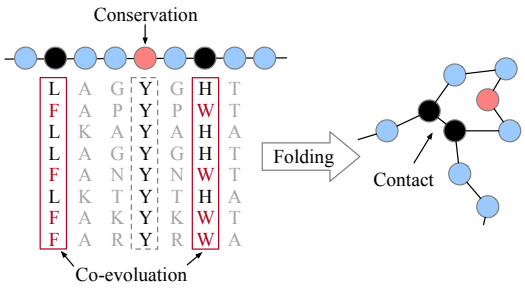
Anonymous Author(s) [1]

Abstract

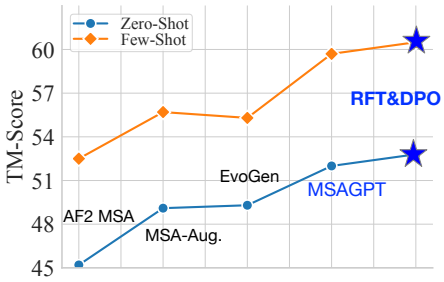
Multiple Sequence Alignment (MSA) plays a pivotal role in unveiling the evolutionary trajectories of protein families. The accuracy of protein structure predictions is often compromised for protein sequences that lack sufficient homologous information to construct high-quality MSA. Although various methods have been proposed to generate virtual MSA under these conditions, they fall short in comprehensively capturing the intricate co-evolutionary patterns within MSA or require guidance from external oracle models. Here we introduce MSAGPT, a novel approach to prompt protein structure predictions via MSA generative pre-training in the low-MSA regime. MSAGPT employs a simple yet effective 2D evolutionary positional encoding scheme to model the complex evolutionary patterns. Endowed by this, its flexible 1D MSA decoding framework facilitates zero- or few-shot learning. Moreover, we demonstrate that leveraging the feedback from AlphaFold2 can further enhance the model's capacity via Rejective Fine-tuning (RFT) and Reinforcement Learning from AF2 Feedback (RLAF). Extensive experiments confirm the efficacy of MSAGPT in generating faithful virtual MSA to enhance the structure prediction accuracy (up to +8.5% TM-Score on few-shot scenarios). The transfer learning capabilities also highlight its great potential for facilitating other protein tasks.

1. Introduction

The advent of deep learning has significantly propelled progress across various scientific domains, exemplified by breakthroughs such as AlphaFold series (Abramson et al., 2024; Jumper et al., 2021) for accurate biomolecular interaction predictions, AlphaGeometry (Trinh et al., 2024) for intricate geometry and mathematical reasoning——to name a few. Among these, AlphaFold2 (AF2) represents a landmark within structural biology, achieving an *in silico* precision of approximately 90% atomic accuracy that rivals wet lab experiments on protein structure predictions (PSP).



(a) A toy example of MSA.



(b) Overall performance comparisons on three natural MSA-scarce benchmark.

Figure 1. **(a) The illustration of MSA and (b) performance comparisons between MSAGPT and advanced baselines.**

The remarkable success of AF2 can be attributed to its innovative use of co-evolutionary information supported by the Multiple Sequence Alignment (MSA). MSA aggregates homologous sequences from vast databases, providing a comprehensive overview of evolutionary trajectories, which is critical for accurately predicting protein structures (Abramson et al., 2024; Jumper et al., 2021; Baek et al., 2021). An illustrative example in Figure 1(a) showcases that the correlations analysis among amino acids (AAs) sites could reveal contacts or conservative regions in the folding structure. Unfortunately, not all proteins possess a rich set of homologous counterparts. Statistical investigations reveal that approximately 20% of metagenomic proteins (Pearson, 2013) and around 11% of proteins from eukaryotic and viral origins (Perdigão et al., 2015) are classified as "orphan" proteins. This presents a significant challenge for MSA-search algorithms in constructing high-quality MSA, consequently

impeding the performance of PSP models (Jumper et al., 2021).

Drawing on the impressive capabilities of large language models endowed either by the autoencoding (Devlin et al., 2018) or the autoregressive language modeling regime (Bubeck et al., 2023; Touvron et al., 2023), protein language models (PLMs) have been developed to unveil the evolutionary patterns and sequence characteristics intrinsic to protein structures. Specifically, generative PLMs (Nijkamp et al., 2023; Chen et al., 2024; Ferruz et al., 2022), trained on vast protein databases (Suzek et al., 2007; Berman et al., 2000; Steinegger et al., 2019b; Steinegger & Söding, 2018) have achieved unparalleled success in generating novel proteins with desired structural properties. These achievements underscore the efficacy of language models in identifying evolutionary patterns within individual protein sequences. Inspired by this, subsequent works (Rao et al., 2021; Zhang et al., 2021) attempt to further integrate MSA as the input or by directly generating virtual yet informative MSA (Truong Jr & Bepler, 2024; Zhang et al., 2023b;a) to provide additional evolutionary insights. These approaches usually adopt customized attentions that merely allow attention aggregated among specific directions, such as axial attention (Ho et al., 2019), for separately analyzing the row- and column-wise co-evolutionary patterns in MSA. However, these attention mechanisms usually have low efficiency in capturing the evolutionary information in MSA, or even fail to adequately capture intricate co-evolutionary dynamics. Taking Figure 1(a) as an example, it is imperative to concurrently analyze the pairwise or high-order relationships of amino acid sites across all homologs to deduce the structural constraints influencing the folding structures, which may not achieved by customized attention. The limited capacity may result in compromised performance on the task that highly resorts to co-evolutionary information.

Built upon the insights mentioned above, we introduce MSAGPT, a simple yet effective framework that prompts protein structure prediction via MSA generative pre-training. This method facilitates *de novo* MSA generation, aiding in protein structure prediction in scenarios with limited MSA available. MSAGPT is characterized by its unique features:

• **2D Evolutionary Positional Encoding.** We employ an innovative dual-axis positional encoding scheme that captures column- and row-wise co-evolutionary information concurrently. This method provides a comprehensive understanding of complex evolutionary relationships with high efficacy. enhancing the model's generative capabilities.

• **1D Zero-/Few-Shot MSA Decoding.** With 2D positional encoding, MSAGPT re-formalizes MSA generation as a one-dimensional sequence generation task, optimized by the simple next-token-prediction objective. This enables MSAGPT to conduct zero- or few-shot MSA generation under a flexible in-context learning framework.

• **Learning from AlphaFold2 Feedback.** MSAGPT further utilizes feedback from AlphaFold2 to reduce hallucinations during MSA generation. This approach ensures the generation of reliable and informative MSA, thus enhancing protein structure prediction.

Extensive experiments conducted on three benchmarks, CAMEO (Haas et al., 2018), CASP, and PDB (Berman et al., 2000), demonstrate the superior capacity of MSAGPT in generating high-quality MSA. Notably, MSAGPT outperforms existing MSA generation models on both zero- and few-shot scenarios. Impressively, AF2 with virtual MSA generated by MSAGPT significantly improves the structure prediction accuracy than that with natural MSA on cases with limited homologous information. Moreover, the subsequent Rejective Fine-tuning (RFT) and learning from AF2 feedback (RLAF) stage further enhance the model's ability to generate informative and faithful MSA, outperforming the original MSAGPT by a large margin, as shown in Figure 1(b). Additionally, we demonstrate that virtual MSA can also benefit other tasks. We expect MSAGPT to become integral in supplementing protein-related tasks requiring critical evolutionary information from MSA. The code, data, and scripts are available here.

2. Related work

Protein Structure Prediction. Proteins are fundamental to the various biological processes that sustain, grow, and protect living organisms. Groundbreaking deep learning approaches, such as AlphaFold series (Abramson et al., 2024; Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021), have been developed to predict the folding structures based on their sequences. These methods have achieved comparable structure prediction accuracy to conventional wet-lab experiments. The success of these cutting-edge methods largely rely on the utilization of MSA, which are retrieved through search algorithms (Zhang et al., 2020; Zheng et al., 2024; Johnson et al., 2010; Steinegger & Söding, 2017) against vast databases (Suzek et al., 2007; Berman et al., 2000; Steinegger et al., 2019b; Steinegger & Söding, 2018). However, challenges arise with "orphan" protein sequences, which lack sufficient homologous sequences for accurate structure prediction. Single-sequence PSP methods (Lin et al., 2023; Chen et al., 2024; Wu et al., 2022; Chowdhury et al., 2022) are designed to infer folding structures directly from the query protein sequences. Despite these advancements, the accuracy of predictions from single-sequence methodologies generally falls short of those derived from MSA-based algorithms.

Protein Language Models. Protein Language Models (PLMs) have emerged as a groundbreaking development

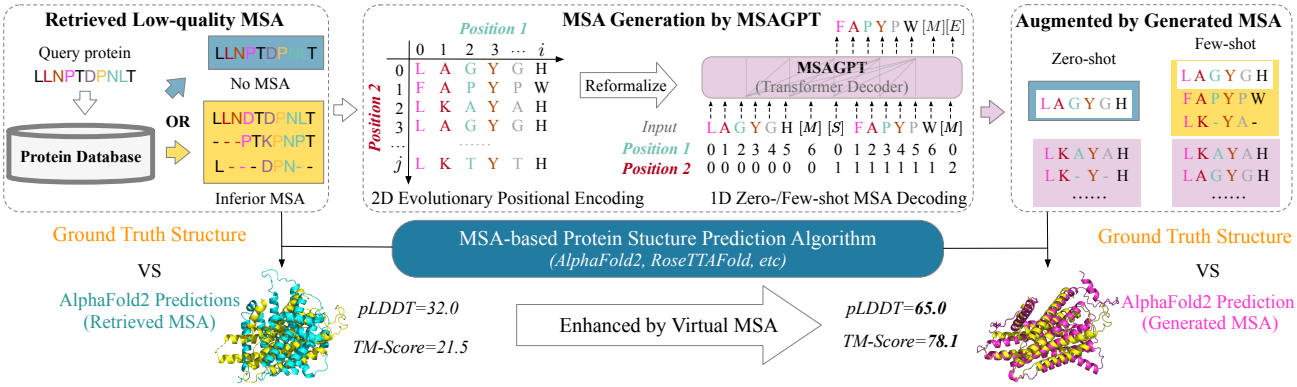


Figure 2. **The overall framework of prompting protein structure predictions via MSA generation. Left**: The challenge faced by conventional search algorithms on protein with scarce homologous sequences, resulting in suboptimal alignments. **Middle-to-Right**: MSAGPT generates informative and high-quality MSA for such challenging queries, presenting a promising approach to overcoming these limitations. `[M]` denotes the sequence separator. `[S]`, `[E]` are the special tokens to represent the start or end of MSA generation.

in computational biology. A family of models including ESM (Lin et al., 2023; Rives et al., 2021), etc (Elnaggar et al., 2021), are trained on single sequences, towards understanding protein structural features. MSA Transformer (Rao et al., 2021) further incorporates MSA as the input, achieving better performance than these single sequence models, underscoring the importance of utilizing the evolutionary information from MSA (Frazer et al., 2021; Chen et al., 2023; Sturmfels et al., 2020). As for protein design, PLMs including ProGen (Nijkamp et al., 2023; Madani et al., 2023) and ProtGPT2 (Ferruz et al., 2022), endowed by the autoregressive training regime, enable the generation of diverse and realistic protein sequences. To enhance MSA generation, MSA-Augmentor (Zhang et al., 2023b) and PoET (Truong Jr & Bepler, 2024) employ the seqs2seqs pre-training, which adopts the sequential axial attention mechanism to capture the evolutionary data across and within the input sequences, both horizontally and vertically. EvoGen (Zhang et al., 2023a), serving as the meta generative model, aims at producing virtual MSA for enhancing protein structure predictions. However, it highly resorts to external structural prediction models to optimize its performance.

2.1. Aligning with Human Preferences

Aligning language models with human preferences has been shown to be effective in improving generation quality (Bubeck et al., 2023; Rafailov et al., 2023; Schulman et al., 2017; Team et al., 2023). Learning from human feedback based on pre-trained models is a common approach in achieving this alignment. Existing methods typically employ supervised fine-tuning using human-annotated datasets or reinforcement learning from human feedback pipelines. Reinforcement algorithms such as Proximal Preference Optimization (Schulman et al., 2017) (PPO) and Direct Preference Optimization (Rafailov et al., 2023) (DPO) are commonly used in these pipelines. Inspired by these, we utilize the feedback from AlphaFold2 to further enhance the generation capability of the pre-trained model, which helps mitigate hallucinations and enables the model to generate accurate and reliable MSA.

3. Preliminary

Definition 3.1. Multiple Sequence Alignment (MSA). Given the query protein sequence $Q \in \mathbb{A}^{1\times L}$, where $\mathbb{A}$ denotes the set of alphabetic symbols used to represent the 20 basic amino acids and L represents the number of amino acids per sequence, the MSA $M \in \mathbb{A}^{N\times L}$ of Q is comprised of N homogeneous protein sequences, which can be obtained either by searching over protein databases or generating with MSA generation methods.

Problem 1. **Prompting Protein Structure Prediction by MSA Generation.** Given Q with initial MSA $M_{\text{init}} \in \mathbb{A}^{n\times L}$ as the prompt, where $n = 0$ indicates the zero-shot MSA generation and $n > 0$ signifies the few-shot MSA generation, we target at learning a function f to generate virtual MSA $M_{\text{gen}} \in \mathbb{A}^{m\times L}$ based on Q and M_{init}, such that the structure prediction accuracy based on the augmented MSA $M_{\text{aug}} \in \mathbb{A}^{(n+m)\times L}$ significantly surpasses that based on the initial MSA M_{init},

$$M_{\text{aug}} = f(Q, M_{\text{init}}),$$
$$\mathbb{I}_{acc}(Q, M_{\text{aug}}) > \mathbb{I}_{acc}(Q, M_{\text{init}})$$

where the $\mathbb{I}_{acc}$ is prediction accuracy comparing the prediction result of AF2 and the ground truth.

In this paper, we mainly focus on improving the structure

prediction accuracy in the low-MSA regime, i.e., the cases that lack a sufficient number of homologous sequences.

4. Methodology

Given a query sequence and its retrieved natural MSA, we aim to comprehensively understand the co-evolutionary patterns in MSA, such that we can generate informative virtual MSA for prompting protein structure prediction in the low-MSA regime. Conceptually, the co-evolutionary information is analogous to the covariance matrix in mathematics, depicting the correlations among amino acids by comparing pairwise or high-order correlations among amino acid sites in MSA, as depicted in Figure 1(a). To achieve this goal, MSAGPT contains two key adoptions, distinguishing it from existing MSA-based PLMs that rely on customized attentions (Jumper et al., 2021; Rao et al., 2021; Zhang et al., 2023b; Truong Jr & Bepler, 2024): **2D Evolutionary Positional Encoding.** Introduces an adaptive dual-axis positional encoding scheme that captures column- and row-wise co-evolutionary information concurrently. And **1D Zero-/Few-Shot MSA Decoding.** Re-formalizes MSA generation as a one-dimensional sequence generation task based on the proposed 2D positional encoding scheme, which enables MSAGPT to conduct zero- or few-shot context learning MSA generation framework. The overall framework is illustrated in Figure 2.

4.1. 2D Evolutionary Positional Encoding

Vanilla transformers typically use 1D positional embeddings to incorporate absolute and relative positional information of tokens. However, when dealing with MSA, which represents stacked homologs, the structure is different. Each row of MSA corresponds to a distinct protein sequence, while each column represents the evolutionary trajectories of a specific amino acids (AAs) site. To effectively capture the evolutionary patterns, recent approaches (Jumper et al., 2021; Rao et al., 2021; Zhang et al., 2023b) have employed decoupled axial attentions, which are designed to capture explicit co-evolutionary information along the rows (protein sequences) and columns (AAs sites). However, these methods often suffer from low efficiency in capturing the information dynamics or fail to capture the evolutionary information adequately.

In light of this, we introduce a novel two-dimensional evolutionary positional encoding scheme, illustrated in Figure 2. Given an MSA $\mathbf{M} \in \mathbb{A}^{N \times L}$, we define a 2D positional id matrix $\mathbf{P} \in \mathbb{N}^{2 \times N \times L}$, where the first positional id $\mathbf{P_0} \in \mathbb{N}^{1 \times N \times L}$ indicates the column position, i.e., $P_0[i, \cdot] = \{0, 1, \cdots, L\}$, and the second positional id $\mathbf{P_1}$ indicates the row position, i.e., $P_1[j, \cdot] = \{0, 1, \cdots, N\}$. The two positional ids are projected into two vectors added to the input token embeddings. We utilize the Rotary Positional Encoding (RoPE) (Su et al., 2024) technique, specifically adapting its two-dimensional variant[1] to suit our 2D positional encoding requirements.

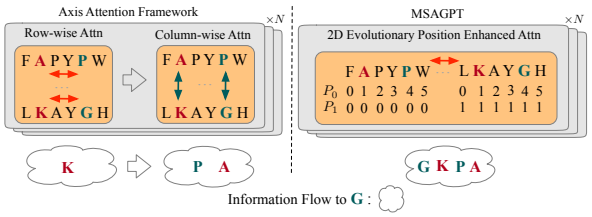


Figure 3. Comparisons among the axial attention (exemplified by (Rao et al., 2021)) and the one in MSAGPT in a single layer. Here we focus on the information aggregated to the AA "G". The 2D evolutionary position enhanced attention shows higher efficiency than the decoupled axial attentions with one-step aggregation to attain sufficient information.

Comparison with Axial Attentions. Considering the 2D positional id (P_0, P_1), the self-attention among AAs (α, β) meets the following unit patterns, as illustrated in Figure 3:

• $P_0^\alpha = P_0^\beta$ & $P_1^\alpha \neq P_1^\beta$. Indicates α and β reside in the same site across different protein sequences, such as the AA pair (A, K) and (P, G), enabling column-wise self-attention that highlights evolutionary patterns conserved across sequences. • $P_0^\alpha \neq P_0^\beta$ & $P_1^\alpha = P_1^\beta$. Suggests α and β are positioned in the same protein sequence but at different sites, such as the AA pair (A, P) and (K, G), facilitating row-wise self-attention that captures sequence-specific features.

• $P_0^\alpha \neq P_0^\beta$ & $P_1^\alpha \neq P_1^\beta$. Denotes α and β lack explicit correlation, such as the AA pair (A, G) and (P, K), may be serving as anchor nodes for complex co-evolutionary information diffusion.

Conceptually, the 2D positional encoding encapsulates the explicit row- and column-wise self-attention patterns with high efficacy. Moreover, it allows unrestricted information diffusion, that is, enabling any two amino acids to attend to one another. Such a framework facilitates unveiling complex co-evolutionary patterns, such as high-order correlations among AAs, that customized self-attentions might overlook.

4.2. 1D Zero-/Few-Shot MSA Decoding

Leveraging the 2D evolutionary positional encoding, we further release the stacked MSA decoding task into the scalable 1D sequence generation framework, without compromising the integrity of co-evolutionary information. Specifically, we convert the MSA $\mathbf{M} \in \mathbb{A}^{N \times L}$ into the flatted 1D sequence $\mathbf{M}^f \in \mathbb{A}^{1 \times NL}$. Similarly, the 2D positional id matrix $\mathbf{P} \in \mathbb{N}^{2 \times N \times L}$ is reshaped into a flattened format,

[1] https://kexue.fm/archives/8397

$\mathbf{P}^f \in \mathbb{N}^{1\times 2\times NL}$. This allows the model to conduct a simple auto-regressive generation process, as illustrated in Figure 2.

Discussions. Admittedly, introducing 2D positional encoding introduces higher time complexity in comparison to conventional customized attention mechanisms (from $O(N^2L) + O(NL^2)$ to $O(N^2L^2)$). However, it is worth noting that the original stacked nature of MSA poses challenges for integrating it with acceleration techniques used in large language models, such as Flash Attention (Dao et al., 2022; Dao, 2023). The 1D decoding framework, conversely, can be easily scaled to accommodate in-context learning frameworks, such as retrieval augmented generation, to further enhance the model's generation capability and expand its application range. From a practical standpoint, the high parallelism of the 1D decoding framework demonstrates superior inference speed, benefiting from techniques like Flash Attention and KV-cache, while incurring negligible memory overhead compared to customized attention mechanisms. For further details, please refer to Appendix Section C.4.

5. The Training Pipeline of MSAGPT

The training pipeline involves three successive stages: **Stage 1: MSA Generative Pre-Training** to obtain the base MSA generation model; **Stage 2: Rejective Fine-tuning (RFT)** to instruct the base model with high-quality MSAs via AF2 annotations, which can reduce generation hallucinations ; **Stage 3: Reinforcement Learning from AlphaFold2 Feedback (RLAF)** to further enhance RFT model's capabilities based on the feedback of AF2. (See Appendix Section C for training details.)

5.1. Stage 1: MSA Generative Pre-Training

Pre-Training Dataset. We utilize the Uniclust30 MSA dataset from OpenProteinSet (Ahdritz et al., 2024), which is processed through an all-against-all search on Uniclust30 (Mirdita et al., 2017) using HHblits (Steinegger et al., 2019a). This results in approximately 16 million MSAs

Pre-training Objective. We adapt the language modeling objective (Radford et al., 2019) to the MSA generation task. The cross-entropy loss for modeling the intrinsic distribution of MSA $\mathbf{M}^f \in \mathbb{A}^{1\times NL}$ is defined as:

$$L_{\text{ce}} = \mathbb{E}_{\mathbf{M}^f}\left[\sum_{i=0}^{N\times L} -\log p(\mathbf{M}_i^f|\mathbf{M}_{<i}^f, \theta)\right] \tag{1}$$

where $\mathbf{M}^f \in \mathbb{A}^{1\times NL}$ is 1D flatted version of the input MSA and θ is the learned parameter.

5.2. Stage 2: Rejective Fine-tuning (RFT)

Noted that the pre-trained dataset inevitably contains noisy co-evolutionary patterns, such as large portions of deletions and insertions, which may mislead the base model to yield hallucinated cases, i.e., the linguistically reasonable but intrinsically unfaithful MSA. Thus we select highly-quality MSAs to further fine-tune the base model via a rejective sampling procedure based on the AF2-annotation.

RFT Dataset. We collect 120,780 protein sequences with structures from Protein Data Bank (PDB) (Berman et al., 2000). For the sequence Q, we search its MSA $M \in \mathbb{A}^{N\times L}$ from UniClust30 (Mirdita et al., 2017) with HHblits (Steinegger et al., 2019a). Then we sample several MSA subsets $\mathbf{m} = \{m_1, m_2, ..., m_i\}$ with replacement, where $m_i \in \mathbb{A}^{n\times L}$ and $n \ll N$. To assure the information density of the sampled data, we filter out the MSA with depth N fewer than $\lceil n \times i/2 \rceil$. Subsequently, we employ AF2 to score the sampled subset using the structure prediction accuracy $\mathbb{I}_{acc}(Q, m_i)$. Then the RFT dataset $\mathcal{D}_{\text{RFT}}$ is defined as:

$$\begin{aligned}\mathcal{D}_{\text{RFT}} = \{(Q, m_i)|(\mathbb{I}_{acc}(Q, m_i)) > \theta_1 \cap \\ (\mathbb{I}_{acc}(Q, m_i) - \mathbb{I}_{acc}(Q, -)) > \theta_2\}\end{aligned} \tag{2}$$

where $\mathbb{I}_{acc}(Q, -)$ indicates the prediction accuracy without using MSAs. Practically, we set the sampling number $i = 10$, the depth of each sampled MSA subset $n = 16$, $\theta_1 = 0.9$, and $\theta_2 = 0.2$, which results in $\mathcal{D}_{\text{RFT}}$ of approximately 60k samples. The base model is fine-tuned on $\mathcal{D}_{\text{RFT}}$ with the same pre-training objective.

5.3. Stage 3: Reinforcement Learning from AlphaFold2 Feedback (RLAF)

We further employ AF2 as the reward model to perform the Reinforcement Learning with AF2 Feedback (RLAF) using Direct Preference Optimization (Rafailov et al., 2023) (DPO) to further guide the RFT model to decode meaningful structure-related MSA patterns that align with the preference of AF2.

RLAF Preference Dataset. For each query Q from the PDB, we use the RFT model to generate its MSA $M \in \mathbb{A}^{N\times L}$ in zero-shot manner. Then, we also sample several MSA subsets $\mathbf{m} = \{m_1, m_2, ..., m_i\}$ and obtain the preference dataset $\mathcal{D}_{\text{DPO}} = \left\{Q^{(k)}, m_{\text{w}}^{(k)}, m_{\text{l}}^{(k)}\right\}_{k=1}^{K}$ as follows,

$$\mathcal{D}_{\text{DPO}} = \{(Q, m_{\text{w}}, m_{\text{l}})|\,(\mathbb{I}_{acc}(Q, m_{\text{w}}) - \mathbb{I}_{acc}(Q, m_{\text{l}})) > \theta_3\} \tag{3}$$

where we set the $\theta_3 = 0.3$, rendering the number of preference data $D_{\text{DPO}} = 11k$.

RLAF Training Objective. The adapted DPO loss is

Table 1. The performance of structure prediction on three benchmarked datasets. avg. Depth represents the average depth of searched MSA across all query sequences. Compared with the base model, the RFT and DPO models achieve higher TM-Score while with lower pLDDT values.

Model	CAMEO (avg. Depth = 8.5)				CASP (avg. Depth = 4.6)				PDB (avg. Depth = 2.6)			
	Zero-Shot		Few-Shot		Zero-Shot		Few-Shot		Zero-Shot		Few-Shot	
	pLDDT	TM	pLDDT	TM	pLDDT	TM	pLDDT	TM	pLDDT	TM	pLDDT	TM
AF2 MSA	63.8	55.4	77.4	71.4	44.0	32.6	54.2	44.1	55.2	45.6	61.0	52.3
MSA-Aug.	67.7	59.2	77.4	72.1	56.8	36.6	63.4	46.3	61.9	49.8	66.0	55.3
EvoGen	66.1	60.3	78.6	75.3	48.2	38.4	55.1	48.5	57.6	49.5	62.8	55.4
MSAGPT	**70.8**	61.4	**80.8**	75.2	**59.0**	39.8	**65.4**	51.0	**68.6**	53.4	**71.3**	59.6
+ **RFT**	68.0	60.5	79.8	76.4	56.8	40.2	64.0	53.6	66.8	53.4	70.3	**60.1**
+ **DPO**	68.9 (+3.1)	**62.7** (+2.4)	80.2 (+2.2)	**76.7** (+1.4)	54.2 (+2.2)	**43.7** (+5.3)	62.7 (+2.0)	**57.0** (+8.5)	64.5 (+6.7)	**53.6** (+3.8)	68.0 (+5.3)	59.7 (+4.7)

defined as:

$$L_{\text{DPO}} = \mathbb{E}_{(Q,m_{\text{w}},m_{\text{l}})\in\mathcal{D}_{\text{DPO}}}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(m_{\text{w}}|Q)}{\pi_{\text{ref}}(m_{\text{w}}|Q)} -\beta\log\frac{\pi_\theta(m_{\text{l}}|Q)}{\pi_{\text{ref}}(m_{\text{l}}|Q)}\right)\right] \quad (4)$$

where π_θ and π_{ref} are initialized by the RFT model and π_{ref} is frozen while π_θ is optimized. During the RLAF training phase, we found that merely using the DPO loss led to training instability. Thus we also adopt the pre-training loss L_{ce} for the chosen answer m_w as a regularization term with the coefficient factor λ in the total loss to mitigate this issue. The total loss $L = L_{\text{DPO}} + \lambda L_{\text{CE}}$, $\lambda = 0.1$. Another critical coefficient β, which measures the penalty intensity of DPO for incorrect answers is set to $\beta = 0.1$.

6. Experiments

6.1. Setup

Benchmarked Dataset. We employ the datasets from CAMEO (Haas et al., 2018), CASP14&15, and PDB (Berman et al., 2000), which are esteemed benchmarks in protein structure analysis spanning a diverse array of biological protein families. For each protein sequence, we search its MSA on UniClust30 database (Mirdita et al., 2017) using HHblits (Steinegger et al., 2019a). Given our focus on addressing the challenge presented by cases with limited MSA information, we build the benchmark to represent the real-world MSA-scarce conditions. More specifically, we identify 200 protein sequences with the number of searched MSA fewer than 20 (8 from cameo, 13 from CASP14&15, 179 from PDB). All MSA of sequences from the test set are removed from the pre-train dataset (see Appendix Figure 9 for details).

Baselines. To assess the performance of MSAGPT, we adopt AF2 as the benchmark MSA-based PSP algorithm. For MSA generation baselines, we compare **MSAGPT**, its RFT-version and its DPO-version with two advanced MSA generation algorithms: **MSA-Augmentor** (Zhang et al., 2023b), which utilizes a sequences-to-sequences pre-training architecture incorporating an encoder and a decoder based on the axial attention (Rao et al., 2021); and **EvoGen** (Zhang et al., 2023a), which employs a meta generative model framework with customized attention, leveraging guidance from AF2 to refine its MSA generation. As PoET (Truong Jr & Bepler, 2024) is designed for mutational scanning tasks, we don't take it as the baseline. Additionally, we include the reference model **AF2 MSA**, which utilizes all the searched natural MSA for prediction.

MSA Generation Pipeline. Given that MSAGPT can perform flexible zero- or few-shot MSA generation to accommodate different levels of available evolutionary information, we define two generation settings to evaluate models' performances under varying conditions:

Zero-Shot Generation. MSA generation is conducted using only the query sequence as input, emphasizing the model's ability to infer necessary evolutionary patterns without additional contexts.

Few-Shot Generation. All the searched natural MSA are viewed as the prompt to inform the few-shot MSA generation process. Then the generated MSA, combined with the initial prompts, serves as augmented data for structure predictions.

Evaluation Metric. We employ TM-Score, a widely-used metric for assessing the structural similarity between predicted structures and ground truth, and pLDDT, a per-residue measure of local confidence, as metrics. All metrics are scaled from 0 to 100 (See Appendix Section D for details).

6.2. MSAGPT's Virtual MSAs Reflect the Co-evolutionary Information

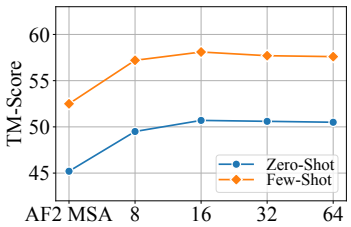
Table 1 showcases the comparative results in three datasets across different baselines. Notably, AF2 MSA, which relies solely on the limited searched MSA without incorporating virtual MSA, exhibits the worst performance. Predictions enhanced with MSA generated by MSA-Augmentor or EvoGen surpass the performance of AF2 MSA. This underscores the critical role of high-quality MSA in enhancing the accuracy of cutting-edge PSP algorithms. Overall, MSAGPT surpasses other advanced baselines by a large margin, achieving +1.4% improvement on CAMEO, +8.5% on CASP, and +4.7% on PDB, as measured by TM-Score. This significant improvement demonstrates not only the superior accuracy and effectiveness of MSAGPT but also its robustness in handling cases with noisy or low-quality MSA. Compared with the base model, the RFT and DPO models achieve higher TM-Score but with lower pLDDT values. This discrepancy might arise from the presence of highly confident (according to pLDDT) but lower-scored decoys (according to TM-Score), as observed in (Zhang et al., 2023a), indicating that aligning with the preference dataset, which is filtered based on TM-Score, makes the model more inclined to generate truly informative MSA rather than hallucinated ones. Statistically, MSAGPT effectively improves the prediction accuracy for 91.0% and 88.9% of protein sequences with limited MSA when compared to AF2 MSA on Zero-Shot and Few-shot scenarios, respectively. This significant finding highlights the potential of our MSAGPT framework to uncover and leverage co-evolutionary patterns within bio-sequences.

6.3. Rethinking the MSA Selection Strategy

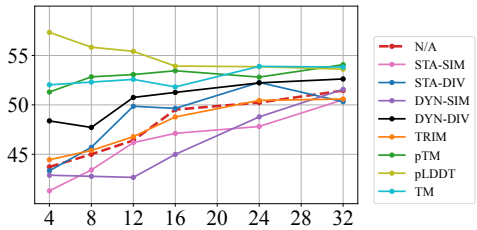
We further study the effect of different depths of virtual MSA, as shown in Figure 4(a). We observe a trend where the relative improvement in structure prediction accuracy decreases as the depth of virtual MSA increases. The accuracy based on MSA with 64 MSA sequences even underperforms those based on only 16 or 32 sequences. We hypothesize that increasing the number of virtual MSA beyond a certain threshold may introduce a dilution effect, where the density of valuable co-evolutionary signals is compromised by the inclusion of the hallucinated generation noise. To alleviate this, we explore MSA selection strategies for filtering out low-quality, noise-inducing sequences while retaining those that contribute positively to the accuracy of structure predictions, as illustrated in Figure 4(b) (See Appendix Section E for details).

1D Sequence Similarity or Diversity Measure. We first arrange MSA by their similarity to the query sequence in descending order. The results reveal that prioritizing MSA based on their high similarity to the query, termed as *static similarity (STA-SIM)*, does not improve prediction accuracy compared to the non-selection approach (N/A). On the contrary, the *static diversity (STA-DIV)* strategy, which favors MSA with lower similarity rankings, slightly outperforms the baseline, highlighting the importance of sequence diversity in enhancing MSA quality. Moreover, we employ the dynamic approach, initially selecting the most (or least) similar MSA to the query sequence and progressively incorporating additional MSA based on their average similarity to the cumulatively selected set, termed as *dynamic similarity (DYN-SIM)* and *dynamic diversity (DYN-DIV)*. The results further confirm the advantage of fostering diversity within MSA rather than selecting only the sequences with high similarities to the query sequence. We also inspect the effectiveness of the widely-adopted MSA *trimming (TRIM)* strategy (Zhang et al., 2023a), which yields a similar TM-Score to the non-selection baseline, undermining its efficacy in selecting MSA with high quality.

3D Structure Affinity Measure. We assume that the generated sequence with high quality should exhibit structural



(a) Effects of MSA Depths



(b) Effects of Selection Methods

Figure 4. The effect of different MSA depths and selection methods. The X-axis indicates the different MSA depths. The Y-axis represents the TM-Score. The dashed line denotes the non-selection baseline.

Table 2. Performance comparison between non-selection and pLDDT-selection models.

Model	CAMEO	CASP	PDB
	TM	TM	TM
MSAGPT-DPO	76.7	57.0	59.7
+ pLDDT Selection	**77.5**	**57.6**	**60.5**

Table 3. Performance comparison between with or without virtual MSA generated by MSAGPT on four protein tasks. ACC is short for Accuracy.

Model	CtP	SsP	LocP	MIB
	ACC	ACC	ACC	ACC
w/o Virtual MSA	11.6	66.5	**58.3**	57.5
w/ Virtual MSA	**13.1**	**69.0**	56.4	**60.3**

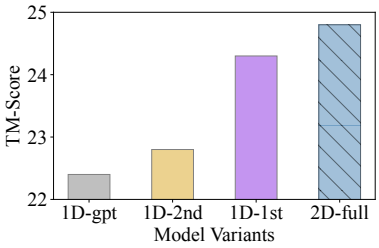


Figure 5. Ablation study with positional encoding variants.

congruity with the query sequence, thereby emitting strong co-evolutionary signals. To validate this, we rank sequences within MSA by their predicted tertiary structures according to the pTM, a predicted TM score (Jumper et al., 2021), pLDDT, and TM-Score, from highest to lowest. These approaches, especially when guided by the pLDDT score, consistently select high-quality MSA, evidenced by the enhanced TM-Score. We compare the non-selection methods (N/A) and pLDDT selection methods on the three benchmarked datasets on few-shot generation scenarios in Table 2. This confirms our hypothesis that structural similarity plays a crucial role in effective MSA selections.

6.4. Transfer Learning of MSAGPT

Since protein structures largely dictate their functions, the virtual MSA, enhancing structure prediction, should similarly benefit other protein tasks. To validate this, we focus on two protein structural tasks: Contact Prediction (CtP) and Secondary Structural Prediction (SsP) and two protein functional tasks: Localization Prediction (LocP) and Metal Ion Binding (MIB) (Chen et al., 2024). We sample 1,000 sequences from each benchmark and conduct 5-fold cross-validation (See Appendix Section D.2 for details). **Results.** Table 3 demonstrate that incorporating MSA from MSAGPT consistently surpasses merely using the single sequence on most tasks. However, it achieves inferior performance on the LocP task, which agrees with the observation (Li et al., 2024) that protein language models may not present scaling behavior on several protein functional or property prediction tasks. Nevertheless, the results show the great potential of MSAGPT to contribute to a wider range of protein-related tasks with generated MSA. We are motivated to explore additional transfer tasks to assess MSAGPT's utility across various domains further.

6.5. Ablation Study

To understand the effect of various positional encoding strategies on capturing co-evolutionary patterns, we design four model variants: **1D_gpt**: Adopts the standard GPT positional encoding; **1D_2nd**: Utilizes only the second-dimensional of the 2D evolutionary positional encoding mechanism; **1D_1st**: Utilizes the first-dimensional positional encoding; **2D_full**: Implements the 2D evolutionary positional encoding mechanism (See Appendix Section D for details).

Results. Figure 5 showcases the TM-score distribution across different model variants. The 1D_gpt exhibits the lowest performance, attributed to its simplistic approach of treating the MSA as a concatenation of homologous sequences, thereby failing to discern any co-evolutionary patterns. Both the 1D_1st and 1D_2nd demonstrate significant improvement over 1D_gpt, by explicitly encoding column- or row-wise relationships within the MSA, respectively. Notably, the performance of 1D_1st is better than that of 1D_2nd, suggesting that column-wise covariance patterns play a more crucial role in structural predictions than row-wise patterns. This aligns with the understanding that the permutation of sequence order does not alter the covariance information among residue sites (Rao et al., 2021). Remarkably, the 2D_full variant, which incorporates the proposed 2D evolutionary positional encoding, outperforms all other models, which underscores its effectiveness in capturing the intricate evolutionary information present in MSA.

7. Conclusion

This paper introduces MSAGPT, a novel approach that prompts protein structure prediction via MSA generative pre-training, to enable accurate protein structure predictions in situations where co-evolutionary information is scarce. To meticulously characterize the co-evolutionary patterns within MSA, MSAGPT designs two innovative techniques: the 2D Evolutionary Positional Encoding scheme and the 1D Zero-/Few-Shot MSA Decoding mechanisms. The post-alignment learning from AlphaFold2 feedback further enhances the quality of MSA generation. Empirical experiments conducted on a variety of benchmarks have demonstrated MSAGPT's robustness and effectiveness. In the future, we plan to apply MSAGPT to broader areas, particularly for tasks that heavily rely on co-evolutionary information, and investigate the limits claimed in the Appendix Section A.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Ahdritz, G., Bouatta, N., Kadyan, S., Jarosch, L., Berenberg, D., Fisk, I., Watkins, A., Ra, S., Bonneau, R., and AlQuraishi, M. Openproteinset: Training data for structural biology at scale. *Advances in Neural Information Processing Systems*, 36, 2024.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chen, B., Xie, Z., Qiu, J., Ye, Z., Xu, J., and Tang, J. Improved the heterodimer protein complex prediction with protein language models. *Briefings in Bioinformatics*, 24 (4):bbad221, 2023.

Chen, B., Cheng, X., Li, P., Geng, Y.-a., Gong, J., Li, S., Bei, Z., Tan, X., Wang, B., Zeng, X., et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.

Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G. M., et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.

Dao, T. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., 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.

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

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., and Schwede, T. Continuous automated model evaluation (cameo) complementing the critical assessment of structure prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86:387–398, 2018.

Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Johnson, L. S., Eddy, S. R., and Portugaly, E. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC bioinformatics*, 11:1–8, 2010.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Li, F.-Z., Amini, A. P., Yue, Y., Yang, K. K., and Lu, A. X. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, pp. 2024–02, 2024.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos Jr, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.

Mirdita, M., Von Den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1):D170–D176, 2017.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. Progen2: exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, 2023.

Pearson, W. R. An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, 42(1):3–1, 2013.

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., et al. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, 112 (52):15898–15903, 2015.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 2021.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Steinegger, M. and Söding, J. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1): 2542, 2018.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics*, 20(1):1–15, 2019a.

Steinegger, M., Mirdita, M., and Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods*, 16(7):603–606, 2019b.

Sturmfels, P., Vig, J., Madani, A., and Rajani, N. F. Profile prediction: An alignment-based pre-training task for protein sequence models. *arXiv preprint arXiv:2012.00195*, 2020.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Truong Jr, T. and Bepler, T. Poet: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36, 2024.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022.

Zhang, C., Zheng, W., Mortuza, S., Li, Y., and Zhang, Y. Deepmsa: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*, 36(7):2105–2112, 2020.

Zhang, H., Ju, F., Zhu, J., He, L., Shao, B., Zheng, N., and Liu, T.-Y. Co-evolution transformer for protein contact prediction. *Advances in Neural Information Processing Systems*, 34:14252–14263, 2021.

Zhang, J., Liu, S., Chen, M., Chu, H., Wang, M., Wang, Z., Yu, J., Ni, N., Yu, F., Chen, D., et al. Unsupervisedly prompting alphafold2 for accurate few-shot protein structure prediction. *Journal of Chemical Theory and Computation*, 19(22):8460–8471, 2023a.

Zhang, L., Chen, J., Shen, T., Li, Y., and Sun, S. Enhancing the protein tertiary structure prediction by multiple sequence alignment generation. *arXiv preprint arXiv:2306.01824*, 2023b.

Zheng, W., Wuyun, Q., Li, Y., Zhang, C., Freddolino, P. L., and Zhang, Y. Improving deep learning protein monomer and complex structure prediction using deepmsa2 with huge metagenomics data. *Nature Methods*, pp. 1–11, 2024.

A. Limitations

In this section, we discuss some limitations that should be resolved in future work.

Scaling behavior of MSAGPT. While we have showcased the effectiveness of MSAGPT in generating informative virtual MSA, it is important to note that our pre-training was conducted with a model containing 2.8 billion parameters. The performance and behavior of MSAGPT , when scaled concerning dataset size, model size, and total compute resources, remain unknown.

General scenarios with sufficient homologs. The primary objective of this paper is to improve the accuracy of protein structure prediction in cases where there is a scarcity of homologous sequences. However, whether we can further enhance the accuracy in scenarios where there are already sufficient MSA available, augmented by virtual MSA, remains an open question.

Transfer Learning on a wide range of tasks. While we have demonstrated the transferability of MSAGPT on several tasks, including protein structure prediction and protein function prediction, its performance on a broader range of tasks remains an open question. The ability of a model to transfer its learned knowledge and adapt to new tasks is a critical aspect of transfer learning. While MSAGPT has shown promising results on the tasks it was evaluated on, it is important to assess its performance on a more diverse set of tasks spanning various domains and problem types.

B. Border Impact

The aim of this paper is to improve the accuracy of protein structure prediction in cases with limited homologous sequences. The generated MSA also shows great potential to transfer to other protein-related tasks. By leveraging the information encoded in the generated MSAs, it is possible to enhance the performance of various protein-related tasks beyond structure prediction. However, the generative MSA may be misused to contaminate the high-quality nature MSA databases. Thus, it is necessary to train a classifier to distinguish the real and MSAGPT-generated MSA according to the intrinsic features.

C. Training Settings and Hyper-parameter Studies.

The overall training pipeline is illustrated in Figure 6.

C.1. Pre-Training

To obtain high-quality MSA, we first screen out clusters with sequence lengths from 25 to 2000, and only retain sequences with the minimum identity of 30% and the largest proportion of gap tokens no more than 10%. The clusters with more than 10 sequences are left. We randomly shuffle the sequences in the MSA to avoid injecting the order bias. Regarding the backbone of MSAGPT, we employ the standard transformer decoder framework (Radford et al., 2019; Brown et al., 2020) and train the model with 2.8 billion parameters owning 36 layers, 2560 embedding size, and 40 attention heads. We employ batches of 48 MSAs with each MSA containing 12,288 residues. We follow BF16 mixed-precision pre-training strategy. We use AdamW (Loshchilov & Hutter, 2018) as our optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, eps $= 10^{-8}$ and a learning rate of 1.2×10^{-4}. We use a cosine learning rate schedule, with a warmup of the first 2.5% steps, and decay the final learning rate down to 10% of the peak learning rate. We use a weight decay of 0.1 and gradient clipping of 1.0 without dropout. For the tokenization of the protein data, we use the residue-level tokenizer which is adopted in several PLMs (Lin et al., 2023; Chen et al., 2024; Nijkamp et al., 2023). To save the GPU memory and accelerate the pre-training process, we substitute the standard self-attention module with the Flash Attention-v1 (Dao et al., 2022) in each layer. All models are trained on 24 A800 GPUs for 254k updates, consuming about 150 billion tokens. This process consumes approximately 150 billion tokens, requiring around 2.7 x 10^{18} floating point operations (FLOPs).

Statistics of Pre-trained Dataset Figure 7 illustrates the length and depth distribution of the pre-training dataset.

C.2. RFT

We fine-tune the base model using the pre-training cross-entropy loss on $\mathcal{D}_{\text{RFT}}$ with training only one epoch. Specifically, we adopt the same experimental settings as that used in the pre-training stage, except for the learning rate of 1.0×10^{-5} by default. Following the pre-training phase, the model undergoes a rejective fine-tuning process, which is more energy-efficient.

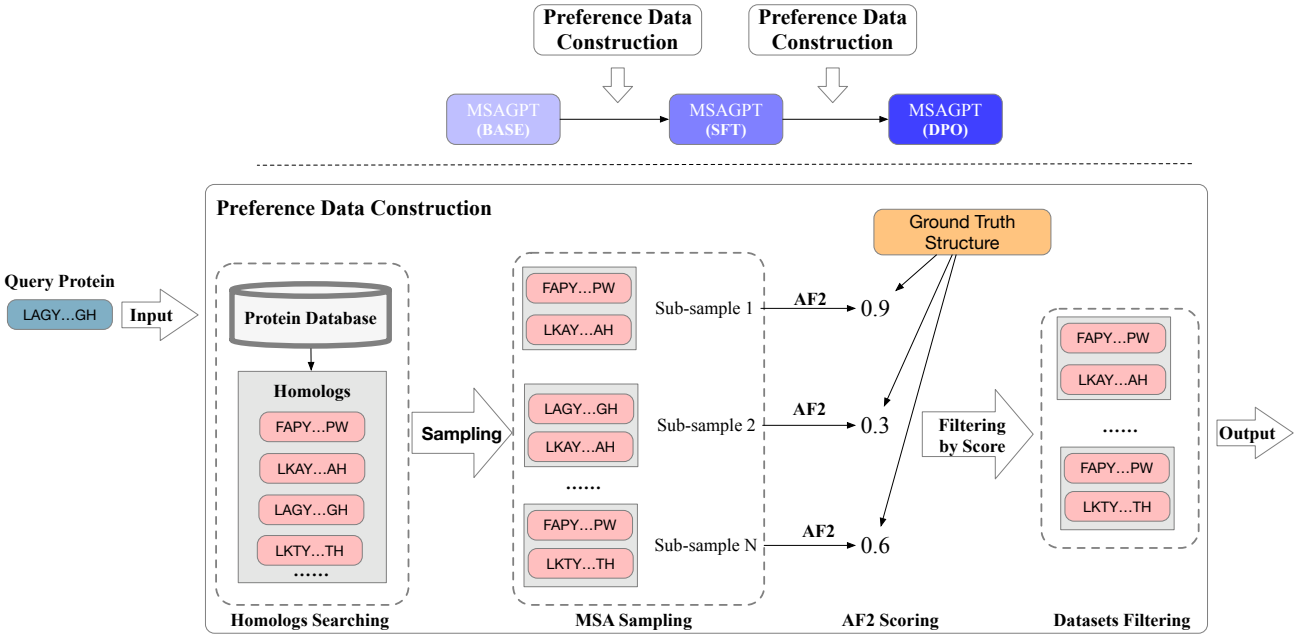


Figure 6. **The overall training pipeline and the illustration of preference dataset construction process for SFT and DPO learning stages.**

Table 4. Performance comparison between different relative improvement threshold θ_2 values.

Threshold θ_2	**0**	**0.2**	**0.5**
MSAGPT+ **RFT**	61.2	**62.5**	61.3

This stage is executed on 8 x A800 GPUs for a single epoch for about two days.

RFT Dataset Filtering Threshold When curating the RFT dataset, we first sample multiple MSA subsets for each protein structure, and select high-quality MSA subsets based on the following criteria: (1) the absolute structure prediction accuracy using the MSA subset, as measured by TM-score, should be larger than θ_1, and (2) the relative improvement of the prediction accuracy after using the MSA subset, as compared to single sequence prediction, should be larger than θ_2. We set $\theta_1 = 0.9$, and experiment with different θ_2 values, as shown in table 4. The RFT model trained with dataset filtered by $\theta_2 = 0.2$ yields the best result, indicating that the relative information gain provided by MSA is a valuable indicator for curating high quality datasets for RFT. Moreover, $\theta_2 = 0.5$ results in a 20% decrease in dataset size, leading to inferior RFT model performance, highlighting the necessity of an intricate balance between data quality and data volumn.

C.3. RLAF

We fine-tune the RFT model using the DPO algorithm on $\mathcal{D}_{\text{DPO}}$ with training only one epoch. Specifically, we adopt the batch size of 1 with each MSA subset containing a maximum of 16,384 residues. We also use AdamW (Loshchilov & Hutter, 2018) with the learning rate of 1.0×10^{-6} by default. We linearly warmup the learning rate from 0 to 1.0×10^{-6} over the first 0.1% steps. This stage is also executed on 8 x A800 GPUs for a single epoch for about one day

Table 5. Performance comparison between different data source and filtering threshold values.

Data Source	**nature(0.2)**	**nature(0.3)**	**generated(0.3)**	**nature(0.3)+generated(0.4)**
MSAGPT+ **RLAF**	62.6	**64.5**	*63.5*	62.7

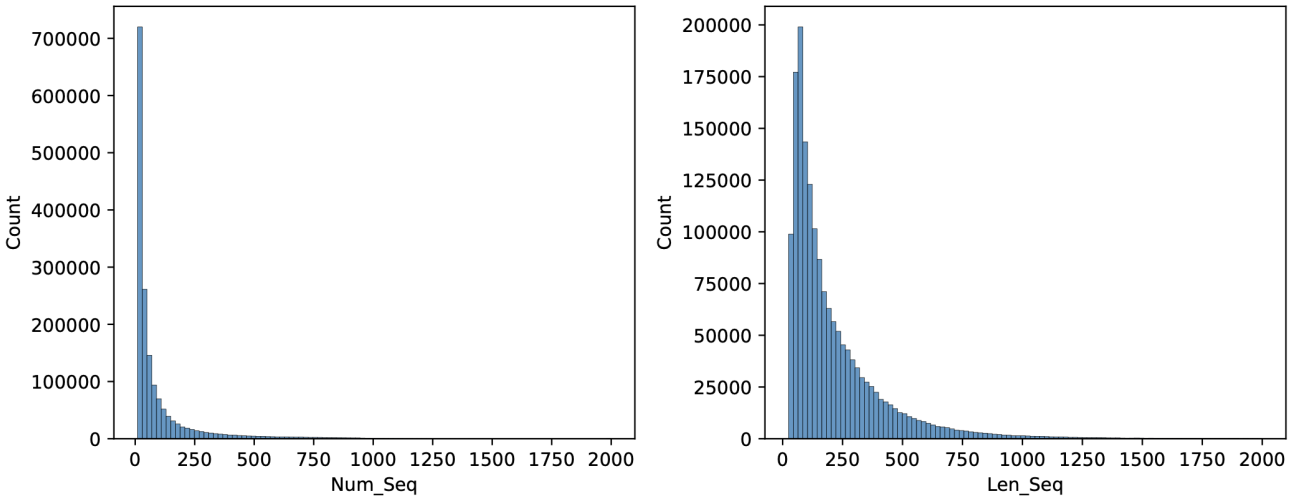


Figure 7. **The length and depth distribution of the pre-training dataset.**

Table 6. The paired Student's t-test between MSAGPT and other baselines on three benchmarks based on the TM-Score, where the p-value less than 0.05 indicates the result is said to be statistically significant.

Model	**CAMEO** (avg. Depth = 8.5)		**CASP** (avg. Depth = 4.6)		**PDB** (avg. Depth = 2.6)	
	Zero-Shot	**Few-Shot**	**Zero-Shot**	**Few-Shot**	**Zero-Shot**	**Few-Shot**
AF2 MSA	0.014	0.023	0.008	0.007	5e-7	8e-9
MSA-Aug.	0.023	0.014	0.044	0.015	6e-6	1e-7
EvoGen	0.038	0.027	0.067	0.016	1e-8	1e-9
MSAGPT	-	-	-	-	-	-

RLAF Dataset. We conducted experiments with different data sources and filtering thresholds θ_3—defined as the minimum relative improvement of the good case over the bad case in DPO data pairs—for the RLAF dataset, as detailed in Table 5. Utilizing only natural MSA subsets sampled from PDB, we found that higher θ_3 values lead to improved model performance, suggesting a correlation between the disparity within data pairs and DPO effectiveness. Interestingly, the quality of MSA subsets generated by the RFT model surpasses that of natural MSA subsets at a θ_3 of 0.2. However, the performance declines when natural MSAs are mixed with generated MSAs, compared to using a single data source during training. This indicates that maintaining distribution homogeneity is crucial for effective DPO alignment.

C.4. Inference Efficiency

Generally, it's vital to consider not just the immediate resource consumption during pre-training and post-alignment, but also the long-term utilization of these models. Once pre-trained, MSAGPT demonstrates significant efficiency, capable of generating protein sequences with up to 100,000 amino acids in under 8 hours. This efficiency underscores the model's value, especially when amortized over its application lifespan and subsequent fine-tunings for specific tasks.

Regarding the scalability of the MSAGPT. We present the inference time with different total lengths (measured by protein sequence length multiply the number of generated sequences.), as shown in Figure 8.

The result showcases MSAGPT's ability to generate substantial sequence lengths within practical time frames, thus affirming its scalability and efficiency.

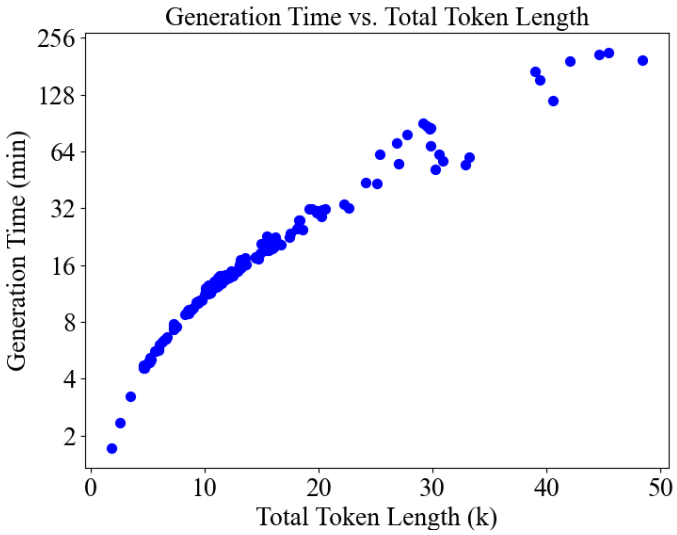


Figure 8. The correlation between total token length (the protein sequence length multiplied by the number of generated MSAs) and the inference time (minutes). In most cases (total token length < 20K), the generation time of MSAGPT is lower than the AF2 search pipeline requiring more than 30 minutes. The result shows MSAGPT can generate substantial sequence lengths within practical time, thus affirming its scalability and efficiency.

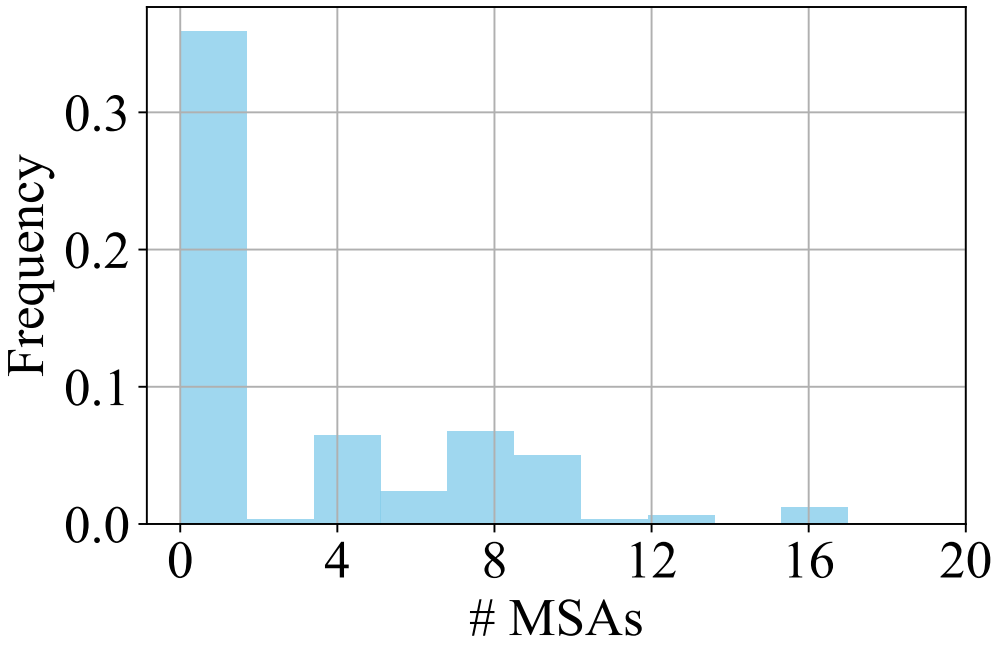


Figure 9. The distribution of MSA depth of benchmarked datasets.

Table 7. The results of 5-fold cross-validation performance between with or without virtual MSA generated by MSAGPT on four protein-related tasks.

Model	1	2	3	4	5	AVG
	Top (L/5)	ACC	ACC	ACC	ACC	-
w/o Virtual MSA (CtP)	11.4	14.3	10.7	9.9	11.8	11.6
w/ Virtual MSA (CtP)	14.1	13.7	13.4	11.8	12.3	**13.1**
w/o Virtual MSA (SsP)	67.7	65.8	64.0	68.9	66.2	66.5
w/ Virtual MSA (SsP)	70.5	67.8	67.5	70.5	69.0	**69.0**
w/o Virtual MSA (LocP)	56.0	64.5	48.0	59.0	57.0	**58.3**
w/ Virtual MSA (LocP)	47.0	58.5	53.5	64.0	59.0	56.4
w/o Virtual MSA (MIB)	58.0	53.5	49.5	59.0	67.5	57.5
w/ Virtual MSA (MIB)	61.5	57.0	63.0	53.0	67.0	**60.3**

D. Experimental Settings

D.1. Evaluation details.

We employ TM-Score, a widely-used metric for assessing the structural similarity between predicted structures and ground truth, and The predicted local distance difference test (pLDDT), a per-residue measure of local confidence. All metrics are scaled from 0 to 100, with higher scores indicating higher confidence and usually a more accurate prediction. where 1 indicates a perfect match between two structures. Each run across 3 independent runs. For each run, we adopt the different temperatures (T $\in \{0.8, 1.0\}$) along with different nucleus sampling factors (P $\in \{0.8, 1.0\}$), experimenting with varying the number of generated MSAs in 8, 16, 32, and 64. The final performance is determined by first identifying the predicted structure with the highest accuracy across these different depths, and then averaging the results across the test set.

D.2. Setup of Transferability of MSAGPT to Other Tasks.

We utilized the MSA Transformer (Rao et al., 2021) as the backbone model with the task-specific heads. We finetune MSA transformer with the head with lr $= 3e-5$ and batchsize $= 16$ on all experiments. All the task benchmarks are obtained following the pipeline in (Chen et al., 2024). For each task, we sample 1000 protein sequences with the corresponding labels. Then we use MSAGPT-DPO to generate 32 virtual MSAs with the generation strategy T=0.8 and P=0.8. Both setups are trained briefly (for one epoch) for 5-fold cross-validation as shown in Table 7, and we report the average performance.

D.3. Setup of Ablation Study

Experiments are conducted based on models with 150 million parameter size encompassing 30 layers, 20 attention heads, 640 embedding dimensions. These models are trained across approximately 30 billion tokens, amounting to around 40k training steps, maintaining consistency in hyper-parameter settings with MSAGPT, except for variations in the positional encoding mechanism. The efficacy of each variant is assessed through zero-shot MSA generation on the CASP test set.

E. Selection Strategy Details and pLDDT Evaluation

To evaluate the effectiveness of different selection strategies, we extracted 4, 8, 12, 16, 24, and 32 sequences from 48 zero-shot generated MSA for each method and computed the median TM-scores (Figure 4(b)) and pLDDT scores (Figure 10) across 33 test cases. The strategies are detailed below.

***Static Similarity / Static Diversity* Strategy:** We select the top-k generated MSA with the highest / lowest sequence identity to the query sequence. Sequence identity is determined by the percentage of identical residues between the two sequences.

***Dynamic Similarity / Static Diversity* Strategy:** Starting with the MSA most / least similar to the query sequence, we sequentially incorporate MSA into the selected set based on the highest or lowest average sequence identity with all sequences already included, until reaching a total of k MSA.

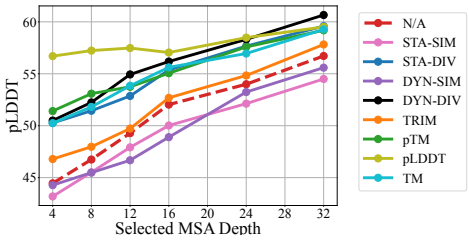


Figure 10. **The pLDDT curves across different selection methods.** Dashed red line represents using all generated sequences of a given depth. Solid lines represent selecting a subset of a given depth from 48 generated sequences with a specific strategy. The curves are smoothed using the Exponential Moving Average with alpha=0.3.

***Trimming* Strategy:** Suggested by EvoGen, this method filters out MSA with less than 50% coverage or sequence identity to the query sequence above 90% or below 20%. Subsequently, it iteratively selects the MSA with the closest sequence identity to the query and an average sequence identity below 90% with all the chosen MSA.

***pTM / pLDDT / TM Score* Strategy:** For each generated MSA, we remove the gaps and predict its structure using AF2. The structures are then ranked based on the pTM score (as reported by AF2), the pLDDT score (as reported by AF2), or the TM score compared to the query sequence's ground truth structure (calculated by US Align), and the MSA corresponding to the top-k structures for each metric are selected accordingly.

F. Protein Structure Prediction Analysis compared with natural MSA

We present a detailed visual comparison of protein structures predicted by AlphaFold2 (AF2) utilizing MSA augmented by MSAGPT, against those predicted with natural MSA. This comparison, as depicted in Figure 12, highlights the remarkable capability of MSAGPTin enhancing the accuracy of structure predictions to levels that closely rival, and in some cases surpass, those based on naturally derived MSA.

We delve into a visualized analysis of protein structures predicted using AlphaFold2 (AF2) with MSA augmented by our proposed model (MSAGPT), alongside those augmented by EvoGen and MSA-Augmenter. This comparison, visualized in Figure 11, encompasses a spectrum of proteins, ranging from short sequences with relatively simple structures, like 7mnv_B, to long sequences with complex configurations, such as 7tdv_B. It includes proteins characterized by multiple beta sheets, exemplified by 7ywg_B, as well as those rich in alpha helices, such as 7tdv_B. Across these diverse cases, MSAGPT significantly surpasses both EvoGen and MSA-Augmenter, improving the TM score to a maximum of 0.9.

By detailed examination, we observe that while the MSA augmented by the baseline models assist AF2 in accurately predicting local structures and folds, they fall short in aligning the global composition and orientation with the ground truth structure, which is effectively addressed by MSA generated by MSAGPT. The local structures, which are generally more discernible from the spatial arrangements of adjacent amino acids, contrast with the global structures whose prediction relies heavily on comprehensively understanding the co-evolutionary information within MSA. These co-evolutionary patterns, indicating proximity in three-dimensional space through simultaneous mutations at multiple positions, are crucial for accurate global structure prediction. These findings underscore MSAGPT's impressive capability to comprehensively capture and utilize co-evolutionary information, thereby significantly enhancing the accuracy of protein structure predictions. More visualization cases about the predictions based on MSA generated by MSAGPT and the predictions based on the natural MSA are illustrated in Appendix F.

G. Protein Structure Prediction Improvement after DPO

Figure 13 represents the comparison before and after the DPO training, depicting notable enhancements in structure prediction accuracy. Figure 14 and 15 provide an in-depth analysis of the generated MSA for each case. Specifically, residues 43, 53, 71-79, 105-111, 122, 132 and 157-166 in the MSA of 7wme_A, along with residues 22-27, 53, and 73 in the MSA of 7sxb_A, display distinct characteristics pre- and post-DPO training.

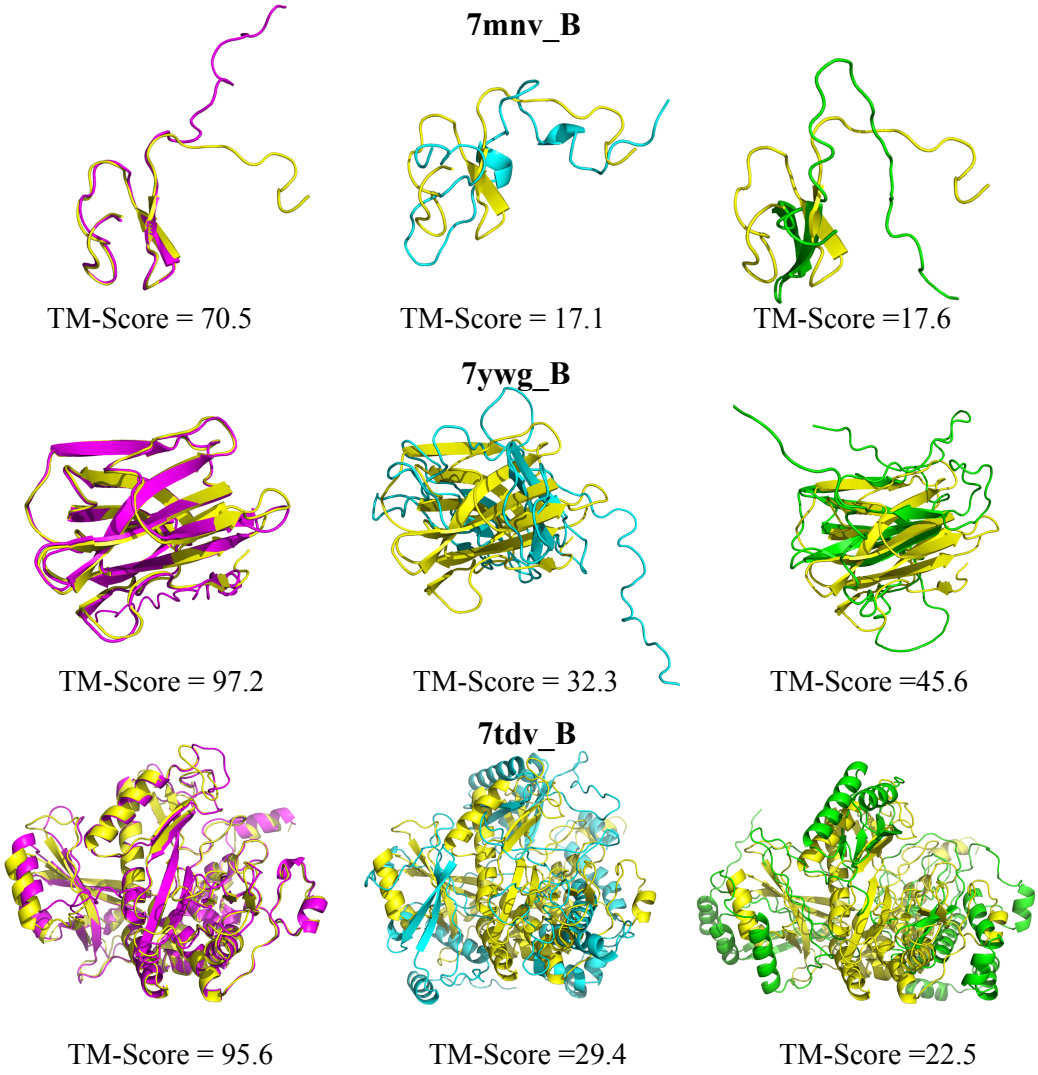


Figure 11. Visualization of improved structure prediction compared with baseline models. Yellow: Ground truth; Pink: Predictions based on MSA generated by MSAGPT; Blue: Predictions from MSA generated by EvoGen; Green: Predictions utilizing MSA generated by MSA-Augmenter.

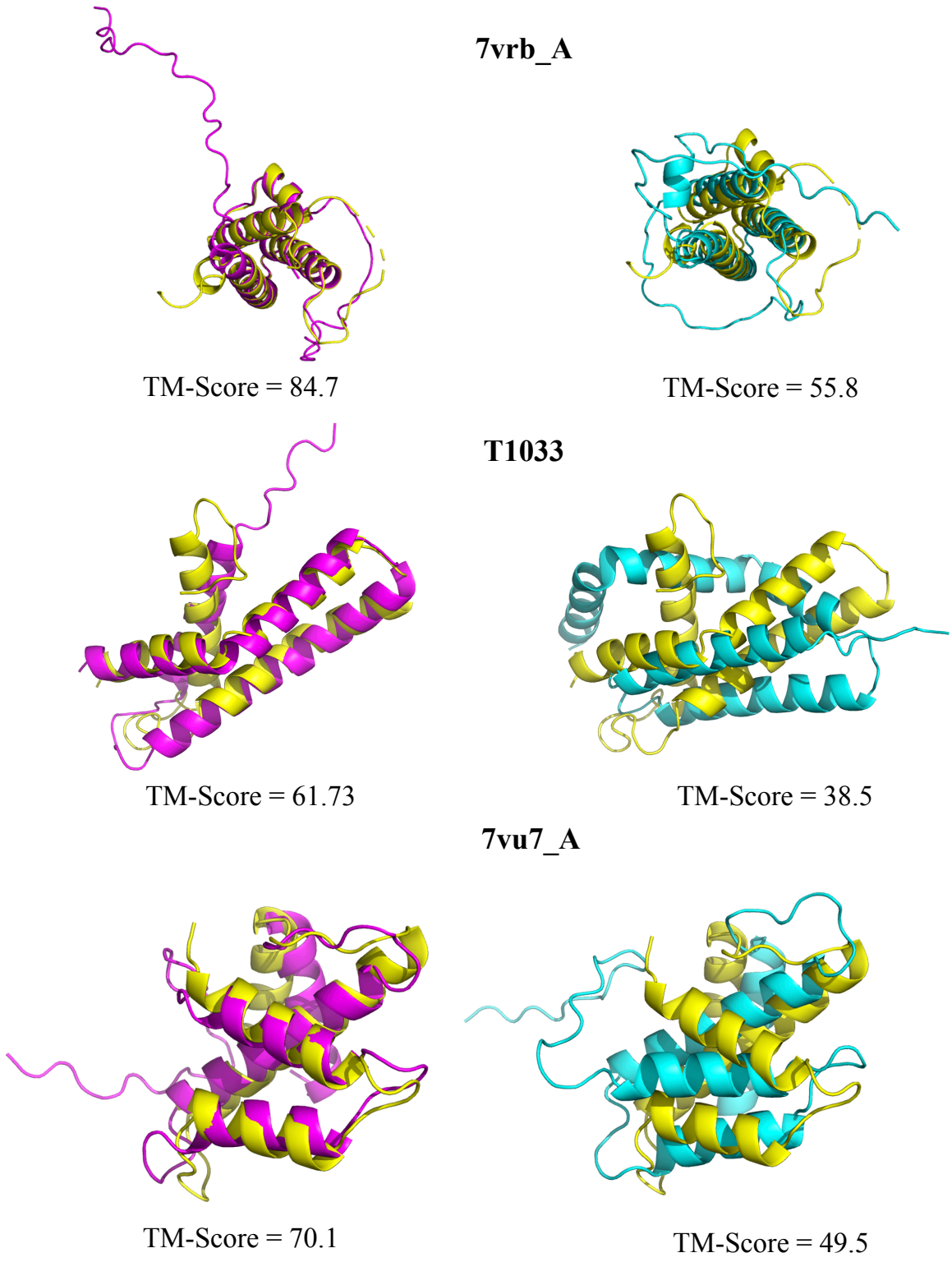


Figure 12. Visualization of improved structure prediction compared with nature MSA Yellow: Ground truth; Pink: Predictions based on MSA generated by MSAGPT; Blue: Predictions from MSA generated by natural MSA.

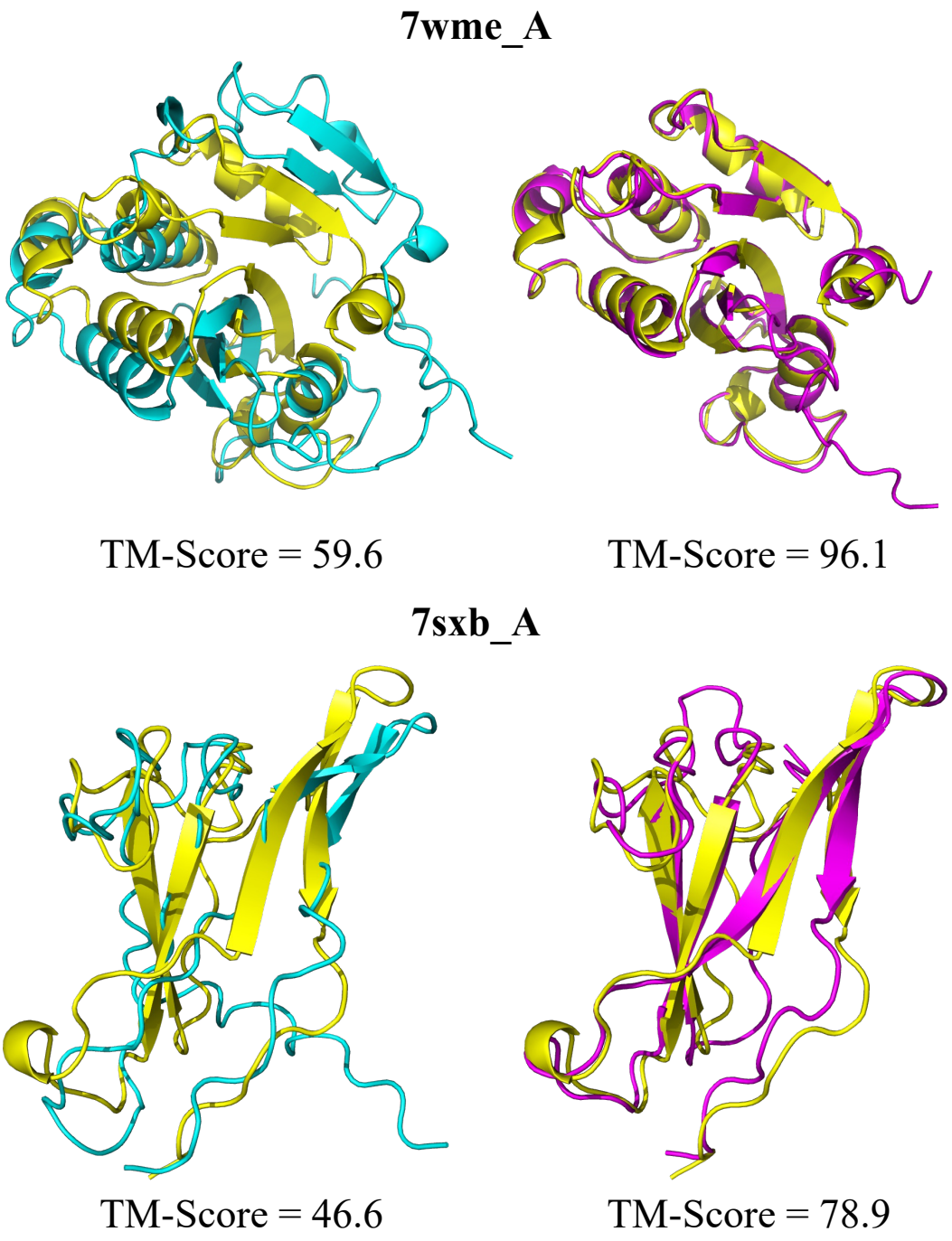
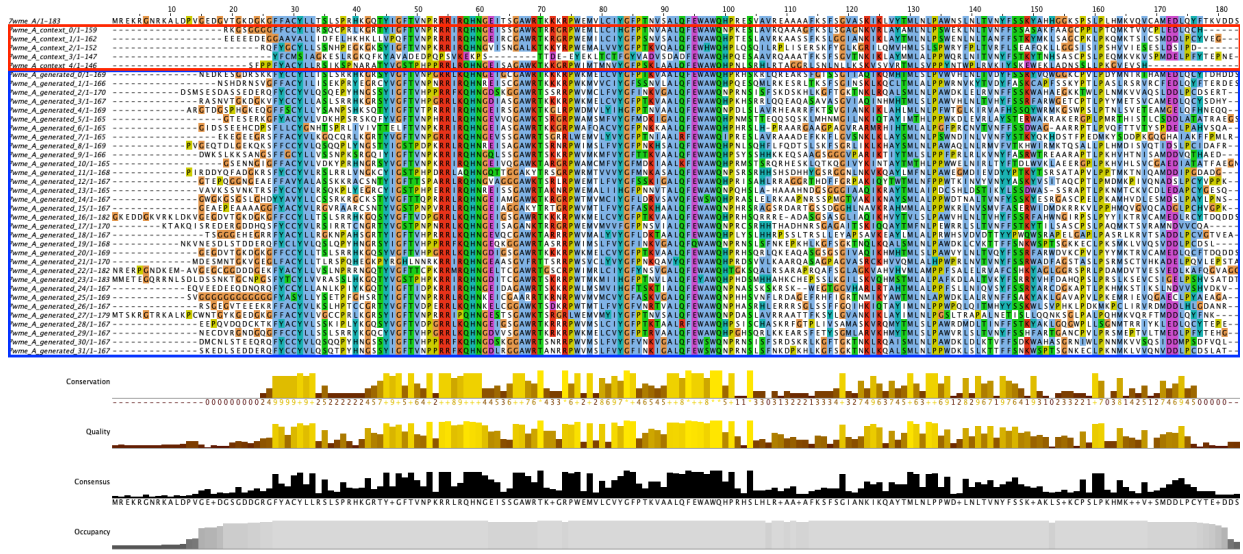
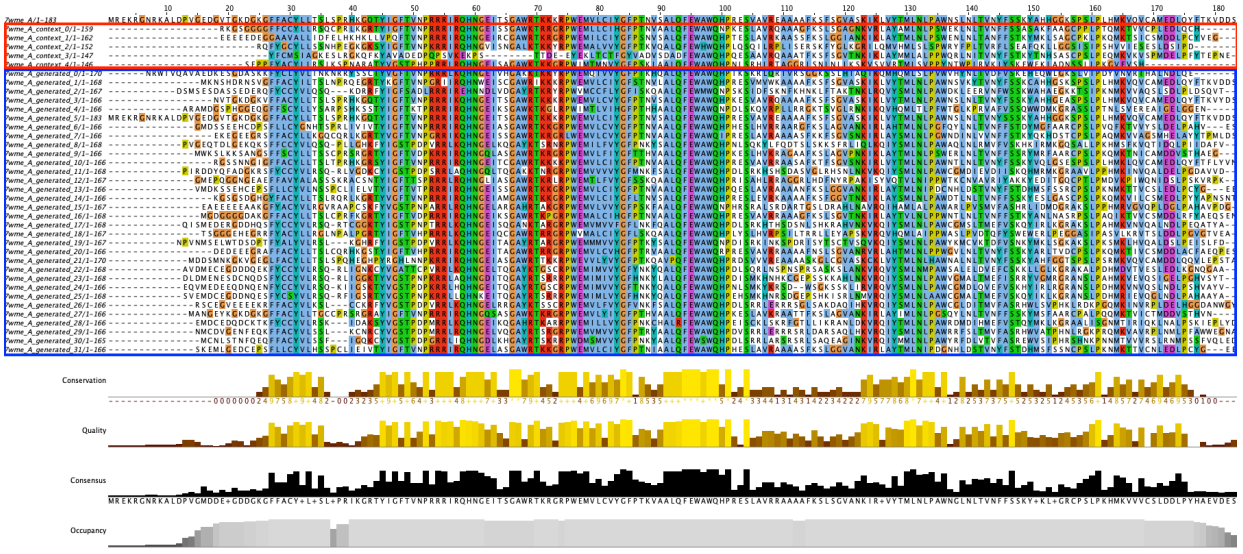


Figure 13. **Visualization of improved structure prediction after DPO.** Yellow: Ground truth; Blue: Predictions based on MSA generated by MSAGPT; Pink: Predictions based on MSA generated by MSAGPT-DPO.;

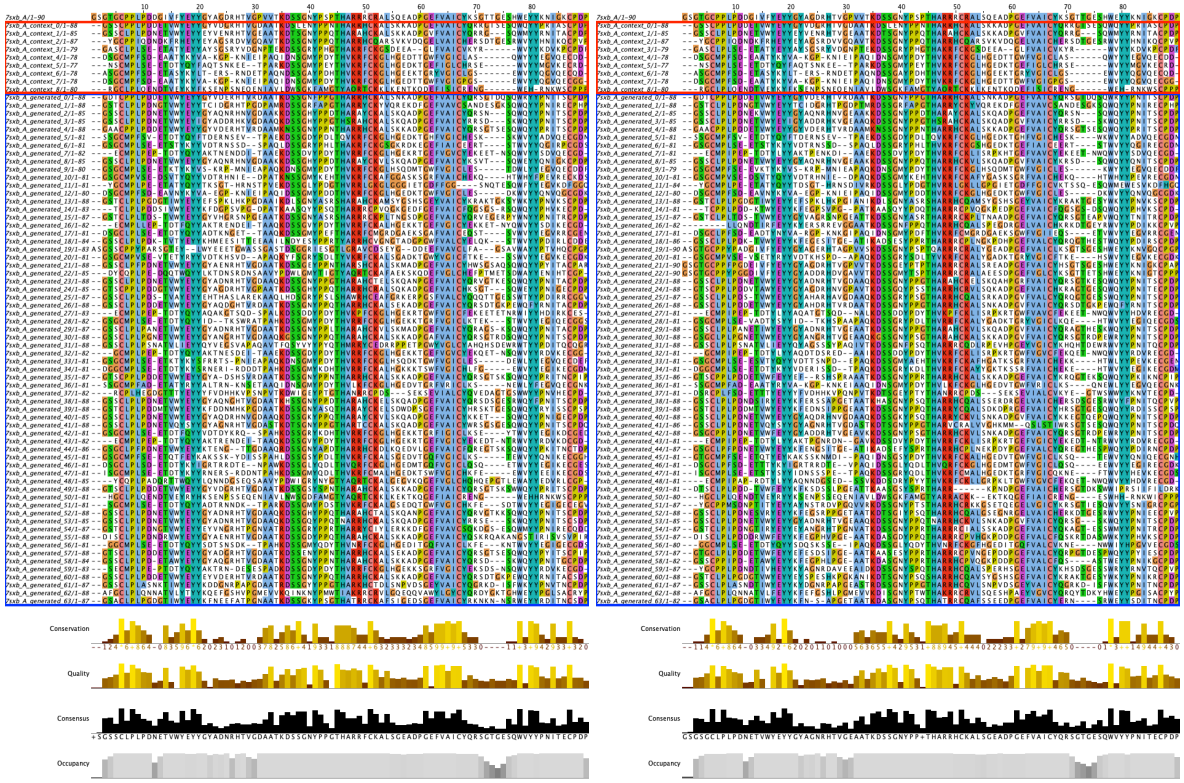


(a) generated by MSAGPT



(b) generated by MSAGPT-DPO

Figure 14. **Residue Distribution of Generated MSA for 7wme_A.** The red box indicates natural MSA used as prompts during generation. The blue box indicates generated MSA. Residues are colored using the clustal scheme by Jalview.



(a) generated by MSAGPT

(b) generated by MSAGPT-DPO

Figure 15. **Residue Distribution of Generated MSA for 7sxb_A.** The red box indicates natural MSA used as prompts during generation. The blue box indicates generated MSA. Residues are colored using the clustal scheme by Jalview.