PLAME: Lightweight MSA Design Advances Protein Folding From Evolutionary Embeddings

Hanqun Cao[1,†] **Xinyi Zhou**[1,†] **Zijun Gao**[1,†] **Chenyu Wang**[2] **Xin Gao**[3]
Zhi Zhang[4] **Chunbin Gu**[1,*] **Ge Liu**[5] **Pheng-Ann Heng**[1]
[1] CUHK [2] MIT [3] UCSD [4] UvA [5] UIUC
[†] Equal contribution. * Corresponding author.

Abstract

Protein structure prediction often hinges on multiple sequence alignments (MSAs), which underperform on low-homology and orphan proteins. We introduce PLAME, a lightweight MSA design framework that leverages evolutionary embeddings from pretrained protein language models to generate MSAs that better support downstream folding. PLAME couples these embeddings with a conservation–diversity loss that balances agreement on conserved positions with coverage of plausible sequence variation. Beyond generation, we develop (i) an MSA selection strategy to filter high-quality candidates and (ii) a sequence-quality metric that is complementary to depth-based measures and predictive of folding gains. On AlphaFold2 low-homology/orphan benchmarks, PLAME delivers state-of-the-art improvements in structure accuracy (e.g., lDDT/TM-score), with consistent gains when paired with AlphaFold3. Ablations isolate the benefits of the selection strategy, and case studies elucidate how MSA characteristics shape AlphaFold confidence and error modes. Finally, we show PLAME functions as a lightweight adapter, enabling ESMFold to approach AlphaFold2-level accuracy while retaining ESMFold-like inference speed. PLAME thus provides a practical path to high-quality folding for proteins lacking strong evolutionary neighbors.

1 Introduction

Understanding complex and variable protein structures is central to target identification, validation, and drug–target interaction studies in drug design [1, 2, 3]. Recent advances such as AlphaFold have transformed structural biology, achieving near-experimental accuracy across a wide range of proteins and complexes [3, 4, 5]. However, most state-of-the-art folding pipelines depend critically on evolutionary information encoded in Multiple Sequence Alignments (MSAs) [3, 5]. As a consequence, their accuracy tracks the quality and depth of the available MSAs, leading to failure modes for low-homology families and orphan proteins where evolutionary neighbors are sparse or absent [6, 7]. In practice, while MSA depth (e.g., effective sequence count) often correlates with predicted confidence (such as pLDDT), this relationship becomes highly unstable in sparse regimes where a few noisy or misaligned sequences can dominate the signal.

Historically, two classes of techniques have been used to mitigate weak homology. Physics-based modeling seeks low-energy conformations under hand-crafted or learned force fields, but is often computationally intensive and limited by approximations in the energy landscape [8, 9]. Template-based approaches leverage homology detection and profile–profile alignment to transfer structural priors from known folds [10, 11], yet degrade in regimes with poor evolutionary signal and are thus ill-suited to orphan proteins. These limitations have motivated data-driven strategies focused on *improving the MSA itself* rather than only the downstream folding network.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Recent MSA design methods fall broadly into two lines (Figure 1). *Sequence-space inpainting* approaches (e.g., MSA Generator, EvoGen) complete or augment partial alignments by learning patterns directly in the discrete sequence space, aiming to reconstruct evolutionary constraints from observed MSAs [12, 13]. *Prompt-based conditional generation* (e.g., MSAGPT, EvoDiff) uses pretrained models conditioned on MSA-style prompts to synthesize additional sequences [14, 15]. These approaches can deepen alignments and improve folding when homologs exist. An orthogonal line of work bypasses explicit MSAs by constructing *implicit* evolutionary representations from single sequences via large protein language models, as in ESMFold [16]. While MSA-free models avoid the homology bottleneck, they also forgo explicit template usage and enhanced homology signals, which can cap ultimate folding accuracy in challenging regimes.

Despite progress, two gaps remain for MSA design aimed at improving folding. **(i) Biased supervision:** Methods trained primarily on extant MSA databases inherit coverage biases toward well-studied families, offering limited guidance for low-homology and orphan proteins. This restricts the learned evolutionary manifold and hampers generalization exactly where better MSAs are most needed. **(ii) Weak alignment–folding linkage:** There is limited understanding—and few lightweight metrics—connecting MSA characteristics to downstream folding outcomes. As a result, generation objectives (e.g., likelihood in sequence space) can be misaligned with what most improves structure accuracy. While fine-tuning folding models with generated data (e.g., MSAGPT) partially closes the loop [14], the compute cost is high and does not yield general, model-agnostic criteria for selecting helpful MSAs. Closing this loop by retraining or adapting a folding model for each candidate MSA is prohibitive at scale, preventing iterative design-and-select cycles across large target sets.

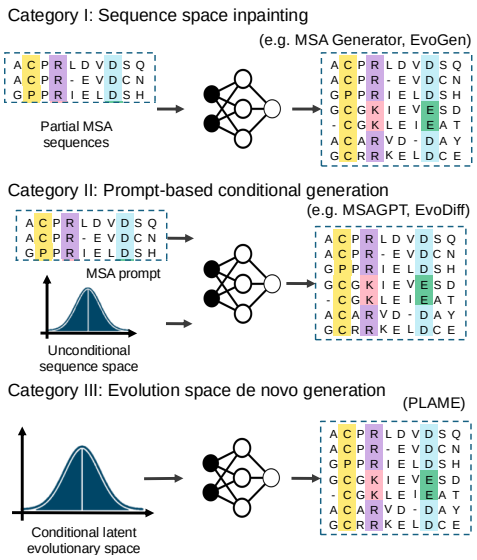


Figure 1: Taxonomy of MSA designers. Prior work models the discrete MSA distribution via sequence inpainting or prompt-based generation. **PLAME** operates in an evolutionary embedding space to *de novo* generate MSAs without requiring sequence prompts.

We introduce **PLAME** (**PL**m **A**ligner for **M**SA **E**nhancement), a lightweight MSA designer targeted at low-homology and orphan proteins. PLAME departs from discrete sequence modeling and instead *generates in an evolutionary embedding space* derived from pretrained protein language models, then decodes to MSAs for downstream folding. This design enables PLAME to synthesize plausible evolutionary neighborhoods even when observed homologs are scarce, while remaining compatible with template-using pipelines.

- **Enriched evolutionary spaces.** PLAME conditions on *evolutionary embeddings* from pretrained protein language models [17, 18, 19], providing richer contextual signals than sparse MSAs for low-homology targets. We introduce a *conservation–diversity loss* that explicitly balances agreement at conserved positions with appropriate sequence variability, steering generation toward biologically plausible and alignment-friendly MSAs.
- **Linking alignment to folding.** We propose a complementary *sequence-quality assessment* metric that captures alignment properties predictive of folding gains, and we operationalize these insights with *HiFiAD* (High-Fidelity Appropriate Diversity), an MSA selection strategy that filters candidates to maximize downstream structure accuracy.
- **Broad, model-agnostic validation.** Across challenging low-homology and orphan datasets, PLAME improves folding accuracy with both AlphaFold2 and AlphaFold3. Ablations isolate the effects of the conservation–diversity loss and HiFiAD on MSA quality and folding outcomes, and case studies spanning general and *de novo* proteins elucidate how specific MSA characteristics modulate AlphaFold confidence and error modes.

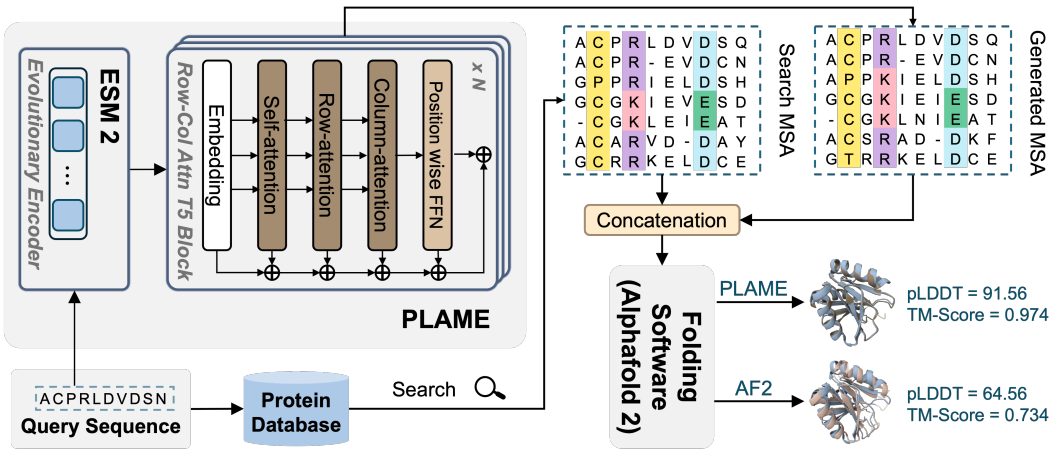


Figure 2: Overview of PLAME framework. PLAME captures ESM-2 evolutionary representations, generating MSAs for augmenting the original MSAs. The augmented MSAs serve as the homology template for folding softwares for folding enhancement. In each block of the T5-architecture, additional row-attention and col-attention are applied to capture co-evolutionary information.

2 Method

2.1 Problem Formulation

Protein structure prediction relies heavily on high-quality MSAs to provide evolutionary information, but the accuracy of folding software $\mathcal{F}_\omega$ significantly drops when MSAs are sparse or insufficient. Given proteins $\mathbf{P} = \{\mathbf{s}, \mathbf{x}, \mathbf{M}\}$, where $\mathbf{s} \in \mathcal{S}$ are query sequences, $\mathbf{x} \in \mathcal{X}$ are 3D structures, and $\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_N\} \in \mathcal{M}$ are MSAs with each $\mathbf{m}_i$ as an aligned homologous sequence. The goal of MSA design models $\mathbf{p}_\theta : \mathcal{M} \to \mathcal{M}$ is designing augmented MSAs $\mathbf{M}_{\text{aug}}$ that enhances evolutionary information to obtain more accurate structures $\mathbf{x}'$ using folding software $\mathcal{F}_\omega$.

$$\mathbf{M}' = \mathbf{p}_\theta(\mathbf{M}), \quad \mathbf{x}' = \mathcal{F}_\omega(\mathbf{s}, \mathbf{M}_{\text{aug}}) \tag{1}$$

where the augmented MSAs are composed of original MSAs $\mathbf{M}$ and generated MSAs $\mathbf{M}'$, denoted as $\mathbf{M}_{\text{aug}} = \{\mathbf{M}, \mathbf{M}'\}$. The quality of the enhanced structures is evaluated using several metrics, including RMSD, TM-score, and pLDDT.

The key to high-fidelity MSA generation lies in constructing an informative evolutionary distribution $\mathbf{z}_{\text{evo}}$, which serves as the foundation for generating augmented MSAs $\mathbf{M}_{\text{aug}}$. Current methods utilize deep neural networks $\mathbf{f}_\theta$ to learn hidden evolutionary distributions directly from existing MSAs.

$$\mathbf{z}_{\text{evo}} = \mathbf{f}_\theta(\mathbf{M}) \tag{2}$$

However, relying solely on sequence-level information from MSAs fails to capture the complete evolutionary landscape, particularly when MSA coverage is sparse or incomplete. To overcome this limitation, we propose an evolutionary space based on evolutionary embeddings derived from pretrained protein language models (PLMs) $\mathbf{g}_\phi$.

$$\mathbf{z}_{\text{evo}} = \mathbf{f}_\theta(\mathbf{g}_\phi(\mathbf{s})) \tag{3}$$

2.2 Model Architecture

PLAME employs an encoder-decoder transformer architecture similar to MSA Transformer [20], with adjustments to the T5 block structure [21]. The encoder and decoder incorporate additional row-wise and column-wise attention mechanisms, designed to better capture evolutionary patterns in MSA data (detailed in Fig 2). Additional mechanisms are introduced as follows.

Row Attention The row attention mechanism models inter-sequence dependencies within the evolutionary space, summarizing sequence relationships across the depth of the input representation. Given an input tensor $\mathbf{H}_{\text{enc}} \in \mathbb{R}^{L \times N \times D \times h}$, where L is the sequence length, N is the number of

MSAs, D is the depth, and h is the hidden dimension, a global representation $\mathbf{H}r$ is computed by averaging the hidden states along the depth dimension:

$$\mathbf{H}_r = \frac{1}{D}\sum_{d=1}^{D}\mathbf{H}_{\text{enc}}^{d}, \quad \mathbf{H}_r \in \mathbb{R}^{L\times N\times h}. \tag{4}$$

Here, $\mathbf{H}_r$ encodes the evolutionary space, facilitating cross-attention during decoding. The cross-row attention (Row-Attn) is defined as:

$$\text{Row-Attn}(\mathbf{Q}_r, \mathbf{K}_r, \mathbf{V}_r) = \text{softmax}\left(\frac{\mathbf{Q}_r\mathbf{K}_r^{\top}}{\sqrt{h}}\right)\mathbf{V}_r. \tag{5}$$

where $\mathbf{Q}_r, \mathbf{K}_r, \mathbf{V}_r$ denote the Query, Key, and Value matrices, respectively.

Column Attention Column attention captures conservation patterns across columns in the MSA, focusing on evolutionary signals at specific positions. To implement this mechanism, the input representation matrix $\mathbf{X}_{\text{dec}} \in \mathbb{R}^{D\times N\times L\times h}$ is first transposed to $\mathbf{X}_{\text{dec}}^{\top} \in \mathbb{R}^{L\times N\times D\times h}$. The self and cross-column attention mechanisms operate on this transformed representation.

For cross-column attention, the query, key, and value matrices $\mathbf{Q}_{\text{c}}, \mathbf{K}_{\text{c}}, \mathbf{V}_{\text{c}}$ and their corresponding projection weights $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are defined as:

$$\mathbf{Q}_{\text{c}} = \mathbf{X}_{\text{dec}}^{\top}\mathbf{W}_q, \quad \mathbf{K}_{\text{c}} = \mathbf{H}_{\text{enc}}^{\top}\mathbf{W}_k, \quad \mathbf{V}_{\text{c}} = \mathbf{H}_{\text{enc}}^{\top}\mathbf{W}_v. \tag{6}$$

The column attention is computed as:

$$\text{Col-Att}(\mathbf{Q}_{\text{c}}, \mathbf{K}_{\text{c}}, \mathbf{V}_{\text{c}}) = \left(\text{softmax}\left(\frac{\mathbf{Q}_{\text{c}}\mathbf{K}_{\text{c}}^{\top}}{\sqrt{h}}\right)\mathbf{V}_{\text{c}}\right)^{\top} \tag{7}$$

Generation & Inference In our implementation, ESM-2 ($\mathbf{g}_\phi$) encodes the query sequence $\mathbf{s}$ into high-dimensional evolutionary embeddings, denoted as $\mathbf{H}_{\text{input}}$. These embeddings are processed by the encoder through N layers of modified T5 blocks, iteratively refining contextualized representations:

$$\mathbf{H}_{\text{Enc}}^{(l)} = \mathbf{Enc}^{(l)}(\mathbf{H}^{(l-1)}), \quad l = 1, \ldots, N, \quad \mathbf{H}^{(0)} = \mathbf{H}_r. \tag{8}$$

The final encoder output, $\mathbf{H}^{(N)}$, captures comprehensive sequence context and is passed to the decoder for autoregressive generation. The decoder generates tokens sequentially, conditioned on the encoder output $\mathbf{H}_{\text{Enc}}^{(N)}$ and previously generated tokens:

$$\mathbf{y}_t = \mathbf{Dec}(\mathbf{y}_{<t}, \mathbf{H}_{\text{Enc}}^{(N)}). \tag{9}$$

The T5 blocks with row and column attention enable the decoder to compute output embeddings, which are then passed through a softmax layer to produce token probabilities for the next position.

2.3 Conservation-Diversity Training Loss

We introduce a position-aware causal inference approach for diverse MSA generation, integrating a PSSM-Weighted Cross-Entropy Loss and a Diversity Regularization Loss to balance focus on conserved regions with sampling diversity.

PSSM-Weighted Cross Entropy (PCE) Loss The PSSM-Weighted Cross Entropy Loss emphasizes accurate predictions in conserved regions of the MSA, which are critical for maintaining protein structure and function. For a single sequence, it is defined as:

$$\mathcal{L}_{\text{seq}} = -\sum_{l=1}^{L} w_l \cdot \log p(y_l \mid y_{<l}), \tag{10}$$

where L is the sequence length, y_l is the true label at position l, and $p(y_l \mid y_{<l})$ is the predicted probability of y_l.

The position-specific weights w_l are derived from the Position-Specific Scoring Matrix (PSSM) and reflect the conservation level at each position. These weights are normalized to the range $[1-\delta, 1+\delta]$, where δ controls sensitivity to conservation. Specifically,

$$w_l = 1 + \delta \cdot \frac{\text{freq}_l - \min(\text{freq})}{\max(\text{freq}) - \min(\text{freq})}. \tag{11}$$

where freq denotes the residue-frequency of 20 types of amino acids. During model training, we apply $\delta = 0.5$, assigning higher weights to conserved positions and lower weights to less conserved ones. For a batch of N sequences, the PCE loss averages over all sequences and positions:

$$\mathcal{L}_{\text{PCE}} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{l=1}^{L_j} w_l^{(j)} \cdot \log p(y_l^{(j)} \mid y_{<l}^{(j)}), \tag{12}$$

where L_j is the length of the j-th sequence, and $w_l^{(j)}$ is the weight for position l in sequence j. This loss emphasizes conserved regions while allowing flexibility in less conserved areas.

Diversity Regularization (DIRE) Loss The Diversity Regularization Loss promotes biological diversity in MSAs by maximizing the entropy of predicted amino acid distributions. For a single sequence, the entropy at position l is calculated as:

$$H_l = -\sum_{a \in \mathcal{A}} p(a \mid y_{<l}) \log p(a \mid y_{<l}), \tag{13}$$

where $p(a \mid y_{<l})$ is the predicted probability of amino acid a at position l, and $\mathcal{A}$ is the set of all amino acids. To encourage diversity across a batch of N sequences, we compute the average entropy over all positions and sequences:

$$\mathcal{L}_{\text{diversity}} = -\frac{1}{N} \sum_{j=1}^{N} \frac{1}{L_j} \sum_{l=1}^{L_j} H_l^{(j)}, \tag{14}$$

where L_j is the length of sequence j, and $H_l^{(j)}$ is the entropy at position l in sequence j. This loss encourages the model to capture the natural amino acid diversity in homologous sequences.

Combined Loss Function The combined loss function balances conservation and diversity:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{PCE}} + (1 - \alpha) \cdot \mathcal{L}_{\text{DIRE}}, \tag{15}$$

where $\alpha \in [0, 1]$ controls the trade-off between the two terms. We set $\alpha = 0.9$ in our experiments, prioritizing conservation while maintaining sufficient diversity. This design ensures the generated MSAs accurately capture conserved regions and retain natural sequence variability. Our theoretical analysis in Appendix A demonstrates that PCE Loss enhances the model's understanding of evolutionary information (MSA profile), while DIRE Loss functions as a regularizer to prevent neglect of variable regions.

2.4 MSA Selection Method – HiFiAD

Existing MSA generation models struggle in ensuring the quality of generated MSAs. Different methods capture MSA distributions with varying effectiveness, and excessive concatenation of virtual MSAs can degrade the quality of the original MSA. This, in turn, restricts the ability of protein structure prediction software to interpret evolutionary distributions. While MSAGPT has explored MSA selection in its ablation studies, existing approaches lack a clear definition of MSA quality and fail to propose systematic selection methods based on sequence-level quality metrics [14].

During MSA augmentation, virtual MSAs inevitably introduce noise, limiting the performance of folding software. A common issue with current models is the generation of over-conserved sequences that closely resemble the query sequence. When these sequences are over-concatenated, they distort the evolutionary distribution of the original MSA. This issue was highlighted in MSAGPT's studies, which showed that performance plateaued once the number of generated MSAs exceeded 32. To address this, balancing recovery rate distributions by selecting sequences with both high and low recovery rates has proven more effective than focusing exclusively on one or the other.

To address these issues, we propose HiFiAD to balance fidelity and diversity for MSA selection while maintaining efficiency. For orphan proteins without original MSAs (termed **zero-shot cases**), we select the top-k sequences with the highest S_{BLOSUM}, along with sequences from the top and bottom $k/2$ of the recovery rate distribution. The BLOSUM substitution score is calculated as:

$$S_{\text{BLOSUM}}(m_i, s) = \sum_{j=1}^{L} B(s_j, m_{ij}), \quad \forall m_i \in M, \tag{16}$$

where B is the BLOSUM matrix (BLOSUM62 in our case); s_j represents the j-th amino acid of sequence S with length L, and m_{ij} refers to the j-th position of the i-th MSA sequence with length L. For proteins with existing MSAs (termed **few-shot cases**), we limit the total number of concatenated MSAs to reduce noise in the original MSA distribution. Specifically, we cap the number of MSA sequences $N_{\text{aug}_{\text{max}}}$ at $\max(16, 2N)$. If fewer than $N_{\text{aug}_{\text{max}}}$ virtual MSAs are available, we retain all generated MSAs to ensure sufficient diversity for downstream analysis.

Table 1: Performance metrics across different modes and models. The best results in each folding mode are highlighted in bold. Zero and Few indicate zero-shot (proteins without MSAs) and few-shot cases (proteins with existing MSAs), respectively.

	pLDDT		GDT		TMscore		RMSD		LDDT		pTM	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few
ESMFold	66.26	62.62	0.6	0.53	0.6	0.57	9.58	12.04	0.62	0.59	/	/
Mode1												
AF2 MSA	60.07	62.14	0.50	0.52	0.50	0.57	12.34	12.16	0.54	0.58	0.44	0.49
EvoDiff	58.68	61.83	0.46	0.50	0.46	0.54	13.81	12.95	0.50	0.56	0.40	0.48
MSAGPT	59.81	61.18	0.48	0.51	0.48	0.56	12.62	12.35	0.53	0.57	0.43	0.48
PLAME	**66.54**	**66.08**	**0.53**	**0.54**	**0.53**	**0.58**	**11.48**	**12.14**	**0.57**	**0.60**	**0.49**	**0.52**
Mode2												
AF2 MSA	66.56	66.32	0.51	0.55	0.52	0.60	**12.06**	11.84	0.55	0.61	/	/
EvoDiff	61.98	65.83	0.48	0.53	0.48	0.58	14.23	**11.82**	0.52	0.59	/	/
MSAGPT	64.88	65.96	0.51	**0.56**	0.51	0.60	12.60	11.90	0.55	0.61	/	/
PLAME	**67.77**	**67.48**	**0.53**	0.55	**0.54**	**0.60**	12.62	11.90	**0.57**	**0.61**	/	/
Mode3												
AF2 MSA	70.31	69.61	0.57	0.60	0.57	0.64	**10.53**	**10.24**	0.60	**0.65**	/	/
EvoDiff	64.39	68.54	0.51	0.57	0.51	0.61	13.20	10.81	0.54	0.62	/	/
MSAGPT	68.39	69.30	0.57	**0.60**	0.56	0.64	11.05	10.40	0.59	0.64	/	/
PLAME	**71.50**	**70.48**	**0.58**	0.59	**0.58**	**0.64**	11.41	10.62	**0.60**	0.64	/	/

3 Experiment

Baselines To evaluate PLAME's capability in generating high-fidelity and diverse MSAs, we compared it with several state-of-the-art AI-based MSA generation methods and AlphaFold2's MSA pipeline [3]. The baselines include AF2 MSA, and open-source methods including EvoDiff and MSAGPT [14, 15]. Additionally, to assess the potential advantages of leveraging AlphaFold2's knowledge base in protein structure prediction, we included ESMFold [16], an MSA-free folding software, as reference.

Evaluation Metric We apply the following metrics for structure assessment: 1) local metrics including pLDDT and LDDT; 2) global metrics including GDT, TM-Score, pTM, and RMSD. To comprehensively evaluate the effectiveness of augmented MSAs, we tested three AF2 folding mode, progressing from computationally efficient, basic settings to more comprehensive configurations: 1) **Mode1**: Folding using pTM-3 model without using templates; 2) **Mode2**: Folding using the default 5 models without using templates; 3) **Mode3**: Folding using the default 5 models with templates. To evaluate the quality of generated multiple sequence alignments (MSAs), we employ four sequence-based metrics to quantify alignment fidelity and diversity: **1) Conservation Score** measures the degree of residue conservation at each alignment position. It is defined as $C_i = \frac{\text{Freq}_{\text{max}}(i)}{N}$, where

$\text{Freq}_{\text{max}}(i)$ is the frequency of the most common residue at position i, and N is the total number of sequences. A high score indicates strong conservation, often reflecting functional or evolutionary constraints, while a low score suggests positional variability. **2) Gap Proportion** quantifies the fraction of gaps at each alignment position, calculated as $G_i = \frac{G(i)}{N}$, where $G(i)$ is the number of gaps at position i. Lower values indicate better alignment quality with fewer missing data or alignment errors. **3) Substitution Compatibility** assesses the evolutionary plausibility of aligned residues using BLOSUM62 substitution scores S_{BLOSUM} (Eq. 16). Higher scores indicate evolutionary compatibility and alignment biological relevance. **4) Alignment Entropy** captures residue diversity at each position using Shannon entropy:

$$H_i = -\sum_{r \in \{R_i\}} p(r) \log_2 p(r),$$

where $\{R_i\}$ is the set of unique residues at position i, and $p(r) = \frac{\text{count}(r)}{N}$ is the relative frequency of residue r. Higher entropy signifies greater positional diversity, indicating weaker functional constraints, whereas lower entropy denotes stronger conservation, often associated with functional or evolutionary significance.

Datasets For the training dataset, we use the PDB and UniClust30 subsets from the open-source OpenProteinSet as our data source [22]. The pre-searched MSAs from OpenFold training are also included. We only retain data with at least 64 MSA sequences. To avoid overlap with the test cases, we removed sequences with over 90% similarity using MMSeqs based on UniClust30 clustering results [23, 24]. This process yields an initial dataset of 293,979 samples, which are split into training and validation sets with a 90:10 ratio. For the test dataset, we adopt the curated test cases from MSAGPT [14], which consist of 200 protein samples from three benchmarks: CASP14&15, CAMEO [25], and PDB [26]. Any $> 90\%$ redundancy between the test cases and training dataset is eliminated.

3.1 PLAME as High-quality MSA Designer

General folding performance We conducted experiments across three AF2 folding modes and evaluated six folding-related metrics. Results are shown in Table 1. We analyze the quality of MSAs generated by different methods and their performance in various scenarios. **1) Performance Superiority:** PLAME achieved state-of-the-art performance on most metrics across all modes. This demonstrates that PLAME-generated MSAs effectively enhance AF2 folding, particularly in scenarios involving low-homology or orphan proteins, where significant improvements in prediction accuracy were observed. In contrast, EvoDiff and MSAGPT struggled to produce high-quality MSAs under these conditions. By leveraging the evolutionary latent space provided by ESM-2, PLAME generated biologically meaningful virtual MSAs, outperforming baselines on key metrics such as pLDDT and RMSD. **2) Diminishing Returns:** As more advanced folding configurations were employed (e.g., with more powerful AF2 models or the inclusion of templates), the benefits of MSA augmentation gradually decreased. In Mode3, where templates were used, certain metrics exhibit performance decrease. This is likely because template information already captures much of the necessary sequence and evolutionary information, reducing the marginal utility of additional MSA augmentation. Furthermore, the noise introduced by low-quality MSAs can be amplified in stronger baselines, further impacting performance. **3) Metric Discrepancies:** The enhancement effects across different metrics were not entirely consistent. For instance, the trends for pLDDT and RMSD diverged. While pLDDT, as a confidence metric, reflects the model's confidence in its predictions, RMSD measures the global deviation of the predicted structure from the ground truth. MSA augmentation improved local prediction quality (leading to higher pLDDT), but had limited impact on reducing global structural deviation (resulting in minimal RMSD improvement). Nevertheless, pLDDT alone is sufficient to demonstrate the high quality of the generated MSAs. **4) Noise Issues:** EvoDiff and MSAGPT produced MSAs that consistently underperformed compared to the original AF2 MSAs across all modes. This was particularly evident in low-homology protein scenarios, where the generated MSAs failed to accurately capture evolutionary distributions and often included irrelevant sequences. When concatenated with the original AF2 MSAs, these low-quality sequences introduced additional noise, disrupting the model's inference. This aligns with the hypothesis presented in Section 3.4, where low-quality MSAs were predicted to introduce noise, ultimately impairing folding performance. **5) Few/Zero-shot Consistency:** In Few-shot scenarios, EvoDiff and MSAGPT performed relatively better due to the presence of initial MSAs that provided additional homology information to guide the generation process. In contrast, PLAME does not

rely on initial MSA searches and directly generates high-quality MSAs through its evolutionary latent space. This enables PLAME to excel in Zero-shot scenarios while maintaining consistent performance in Few-shot cases. This advantage significantly reduces runtime costs while ensuring robust enhancements across all scenarios.

3.2 PLAME as MSA-free ESMFold Adapter to AlphaFold2

The comparative analysis between PLAME and ESM-Fold demonstrates PLAME's growing advantage as the folding modes become more advanced. In Mode1 and Mode2, PLAME outperforms ESMFold in key metrics such as pLDDT and RMSD, with obvious gains in pLDDT. In Mode3, the introduction of structural templates amplifies PLAME's advantage, achieving better results across all metrics. For instance, pLDDT of PLAME increases from 66.54 to 71.50, surpassing ESMFold's 66.26. These results validate PLAME's ability to merge ESMFold's computational efficiency with AF2's performance by generating high-quality virtual MSAs. Acting as an adaptive network, PLAME transforms ESMFold's single-sequence inference into AF2's MSA-based folding, integrating evolutionary information to enhance prediction accuracy. Thus, PLAME overcomes ESMFold's limitations by introducing virtual MSA generation, achieving significant improvements in Mode3 and providing an efficient and accurate solution for protein structure prediction.

Table 2: Ablation study over different MSA selection approaches.

	pLDDT		GDT		TMscore		RMSD		LDDT		pTM	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few
Random-16	63.61	62.63	0.52	0.51	0.52	0.56	12.01	12.67	0.55	0.58	0.46	0.49
Blosum-8	61.04	62.71	0.5	0.52	0.51	0.57	12.53	12.69	0.55	0.58	0.45	0.50
Blosum-32	62.97	62.40	0.50	0.50	0.51	0.55	12.28	12.84	0.55	0.57	0.45	0.48
Top-Rec-16	62.04	62.93	0.51	0.51	0.51	0.55	12.15	12.48	0.55	0.57	0.45	0.49
Top-down-Rec-16	63.43	63.10	0.52	0.52	0.51	0.57	11.97	12.15	0.55	0.58	0.46	0.49
HiFiAD	66.54	66.08	0.53	0.54	0.53	0.58	11.48	12.14	0.57	0.60	0.49	0.52

3.3 Ablation on Selection Methods

To evaluate the impact of MSA quality (2.4), we perform an ablation study using different MSA selection strategies. The results are summarized in Table 2. The selection methods are based on sequence similarity metrics (*Rec*) and BLOSUM substitution scores (*BLOSUM*). Additionally, *Top* and *Down* refer to the highest- and lowest-scoring cases. The experiments lead to the following key observations. **1) Effective HiFiAD filtering**: HiFiAD achieves superior performance across all metrics compared to other filtering methods, demonstrating that the combination of high fidelity and balanced diversity is essential. **2) High-fidelity**: Compared to the Top-down-Rec method and random selection, HiFiAD identifies more high-fidelity cases, validating the importance of leveraging in-distribution MSAs. **3) Diverse samples**: Compared to BLOSUM-based methods and Top-Rec, HiFiAD selects more diverse samples. This balanced diversity prevents the conservation distribution from being overly deterministic, which could potentially correct the augmented MSA distribution. Among sequence-metric-based filtering methods, HiFiAD dynamically adjusts to Few- and Zero-shot scenarios, as well as varying MSA quality levels, making it a simple yet effective approach for MSA selection. More ablation results are shown in Table 9 in Appendix.

4 Conclusion

We introduce PLAME, the first model to leverage evolutionary embeddings for MSA sequence generation. PLAME bridges single-sequence inference and MSA-based methods, improving protein folding performance. PLAME-generated MSAs outperform existing methods in conservation and diversity metrics, achieving significant improvements in structure prediction accuracy across protein families. PLAME serves as both an MSA augmenter and an AlphaFold adapter, eliminating MSA searches while providing fast, accurate, scalable structure prediction. Our quality metrics and experiments reveal the relationship between MSA characteristics and folding performance, clarifying how sequence information translates to structural accuracy.

References

[1] David Baker and Andrej Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, 2001.

[2] George A Khoury, James Smadbeck, Chris A Kieslich, and Christodoulos A Floudas. Protein folding and de novo protein design for biotechnological applications. *Trends in biotechnology*, 32(2):99–109, 2014.

[3] 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.

[4] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, et al. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, pages 1–11, 2024.

[5] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[6] Sohee Kwon, Jonghun Won, Andriy Kryshtafovych, and Chaok Seok. Assessment of protein model structure accuracy estimation in casp14: Old and new challenges. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1940–1948, 2021.

[7] Benjamin Webb and Andrej Sali. Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54(1):5–6, 2016.

[8] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. *Methods in Enzymology*, 383:66–93, 2004.

[9] William D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[10] Andrea Hildebrand, Michael Remmert, Andreas Biegert, and Johannes Söding. Fast and accurate automatic structure prediction with hhpred. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):128–132, 2009.

[11] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29–W37, 2011.

[12] Le Zhang, Jiayang Chen, Tao Shen, Yu Li, and Siqi Sun. Enhancing the protein tertiary structure prediction by multiple sequence alignment generation. *arXiv preprint arXiv:2306.01824*, 2023.

[13] Jun Zhang, Sirui Liu, Mengyun Chen, Haotian Chu, Min Wang, Zidong Wang, Jialiang Yu, Ningxi Ni, Fan Yu, Diqing Chen, et al. Few-shot learning of accurate folding landscape for protein structure prediction. *arXiv preprint arXiv:2208.09652*, 2022.

[14] Bo Chen, Zhilei Bei, Xingyi Cheng, Pan Li, Jie Tang, and Le Song. Msagpt: Neural prompting protein structure prediction via msa generative pre-training. *arXiv preprint arXiv:2406.05347*, 2024.

[15] Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Bob Strome, Alan Moses, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pages 2023–09, 2023.

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

[17] Xiaoyu Wang, Heqian Zhang, Jiaquan Huang, and Zhiwei Qin. Maape: A modular approach to evolutionary analysis of protein embeddings. *bioRxiv*, pages 2024–11, 2024.

[18] Claire D McWhite, Isabel Armour-Garb, and Mona Singh. Leveraging protein language models for accurate multiple sequence alignments. *Genome Research*, 33(7):1145–1153, 2023.

[19] Liang Hong, Zhihang Hu, Siqi Sun, Xiangru Tang, Jiuming Wang, Qingxiong Tan, Liangzhen Zheng, Sheng Wang, Sheng Xu, Irwin King, et al. Fast, sensitive detection of protein homologs using deep dense retrieval. *Nature Biotechnology*, pages 1–13, 2024.

[20] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

[21] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[22] Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Lukas Jarosch, Dan Berenberg, Ian Fisk, Andrew Watkins, Stephen Ra, Richard Bonneau, and Mohammed AlQuraishi. Openproteinset: Training data for structural biology at scale. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J Martin, Johannes Söding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1):D170–D176, 2017.

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

[25] Jürgen Haas, Alessandro Barbato, Dario Behringer, Gabriel Studer, Steven Roth, Martino Bertoni, Khaled Mostaguir, Rafal Gumienny, and Torsten Schwede. Continuous automated model evaluation (cameo) complementing the critical assessment of structure prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86:387–398, 2018.

[26] 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.

[27] Bernard R Brooks, Charles L Brooks, Alexander D Mackerell, Lennart Nilsson, Robert J Petrella, Benoıt Roux, Young Won, Georgios Archontis, Claus Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[28] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature structural biology*, 9(9):646–652, 2002.

[29] Peter L Freddolino, Christopher B Harrison, Yanxin Liu, and Klaus Schulten. Challenges in protein-folding simulations. *Nature physics*, 6(10):751–758, 2010.

[30] Vijay S. Pande, Kyle Beauchamp, and Gregory R. Bowman. Everything you wanted to know about markov state models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

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

[32] L Steven Johnson, Sean R Eddy, and Elon Portugaly. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC bioinformatics*, 11:1–8, 2010.

[33] Wei Zheng, Qiqige Wuyun, Yang Li, Chengxin Zhang, P Lydia Freddolino, and Yang Zhang. Improving deep learning protein monomer and complex structure prediction using deepmsa2 with huge metagenomics data. *Nature Methods*, 21(2):279–289, 2024.

[34] Sewon Lee, Gyuri Kim, Eli Levy Karin, Milot Mirdita, Sukhwan Park, Rayan Chikhi, Artem Babaian, Andriy Kryshtafovych, and Martin Steinegger. Petabase-scale homology search for structure prediction. *Cold Spring Harbor perspectives in biology*, 16(5):a041465, 2024.

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

A Proof of Theorem

We provide additional statements to demonstrate the superiority of the Conservation-Diversity Training Loss. Firstly, we demonstrate that the PCE Loss as a conservation-aware weighted loss by position in the perspective of MSA profiles.

Lemma 1. *Let $P(l,a)$ be the empirical amino–acid distribution for residue $a \in \mathcal{A}$, and let $Q_\theta(l,a)$ denote the model distribution at the residue (i.e. the conditional probability $p_\theta(a \mid y_{<l})$ after taking expectation over prefixes). Assign each column a weight $w_l \in [\,1-\delta,\, 1+\delta\,]$ obtained from its conservation score. Then PCE loss directs optimization preferentially toward conserved positions by minimizing a weighted KL divergence and scaling gradient magnitudes in proportion to w_l.*

Proof. For a sufficiently large set of N homologous sequences sampled from P, the expected cross-entropy loss is

$$\mathbb{E}[\mathcal{L}_{\text{CE}}] = -\sum_{l=1}^{L} \sum_{a \in \mathcal{A}} P(l,a)\, \log Q_\theta(l,a). \tag{17}$$

Re-expressing each column term as $-\sum_a P \log Q = H\big(P(l,\cdot)\big) + \text{KL}\big(P(l,\cdot) \| Q_\theta(l,\cdot)\big)$, we obtain

$$\mathbb{E}[\mathcal{L}_{\text{CE}}] = \sum_{l=1}^{L} \text{KL}\big(P(l,\cdot) \| Q_\theta(l,\cdot)\big) + \sum_{l=1}^{L} H\big(P(l,\cdot)\big). \tag{18}$$

For the PCE loss,

$$\mathbb{E}[\mathcal{L}_{\text{PCE}}] = -\sum_{l=1}^{L} w_l \sum_{a \in \mathcal{A}} P(l,a)\, \log Q_\theta(l,a), \tag{19}$$

which can analogously be rewritten as the position-wise weighted KL

$$\mathbb{E}[\mathcal{L}_{\text{PCE}}] = \sum_{l=1}^{L} w_l\ \text{KL}\big(P(l,\cdot) \| Q_\theta(l,\cdot)\big) + \sum_{l=1}^{L} w_l\, H\big(P(l,\cdot)\big). \tag{20}$$

Let θ denote the model parameters. The gradient of the CE loss for column l is

$$\frac{\partial \mathcal{L}_{\text{PCE},l}}{\partial \theta} = -\sum_{a \in \mathcal{A}} P(l,a)\, \frac{1}{Q_\theta(l,a)} \frac{\partial Q_\theta(l,a)}{\partial \theta}. \tag{21}$$

For PCE the gradient is simply scaled by w_l:

$$\frac{\partial \mathcal{L}_{\text{PCE},l}}{\partial \theta} = -w_l \sum_{a \in \mathcal{A}} P(l,a)\, \frac{1}{Q_\theta(l,a)} \frac{\partial Q_\theta(l,a)}{\partial \theta} = w_l\, \frac{\partial \mathcal{L}_{\text{CE},l}}{\partial \theta}. \tag{22}$$

Consequently, in highly conserved columns the gradient magnitude is amplified by $1+\delta$, whereas in variable columns ($w_l \approx 1-\delta$) it is attenuated, focusing optimization effort on conserved regions. □

Based on the understanding of the PCE Loss, we then demonstrate that PCE Loss is expected to capture evolutionary information (MSA profile) with less error–measured by KL-Divergence.

Theorem 1. *Let $P(l,a)$ be the true amino–acid distribution in column l ($l = 1, \ldots, L$) of an MSA and let $Q_\theta(l,a)$ be the distribution produced by a parametrised generative model Q_θ. Denote the column–wise Kullback–Leibler divergence by*

$$\mathit{KL}\big(P(l,\cdot) \,\|\, Q_\theta(l,\cdot)\big) = \sum_{a \in \mathcal{A}} P(l,a)\, \log \frac{P(l,a)}{Q_\theta(l,a)}. \tag{23}$$

Let

$$\theta^\star_{\text{CE}} = \arg\min_\theta \mathcal{L}_{\mathit{CE}}(\theta), \qquad \theta^\star_{\text{PCE}} = \arg\min_\theta \mathcal{L}_{\mathit{PCE}}(\theta). \tag{24}$$

Define the average profile KL divergence

$$D_{KL}^{avg}(\theta) := \frac{1}{L}\sum_{l=1}^{L} KL\big(P(l,\cdot)\,\|\,Q_\theta(l,\cdot)\big). \tag{25}$$

Under the assumption that both optimization problems are solved to global optimality, the model trained with PCE Loss captures the MSA profile with less divergence D_{KL}^{avg}:

$$D_{KL}^{avg}(\theta^\star_{\mathrm{PCE}}) \;\le\; D_{KL}^{avg}(\theta^\star_{\mathrm{CE}}) \tag{26}$$

Proof. Rewrite two losses in the form of KL-Divergence $\sum_a P\log Q \;=\; H\big(P(l,\cdot)\big) + \mathrm{KL}\big(P(l,\cdot)\|Q_\theta(l,\cdot)\big)$, we have:

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}}(\theta) &= C_0 + \sum_{l=1}^{L}\mathrm{KL}\big(P(l,\cdot)\|Q_\theta(l,\cdot)\big),\\ \mathcal{L}_{\mathrm{PCE}}(\theta) &= C_w + \sum_{l=1}^{L} w_l\;\mathrm{KL}\big(P(l,\cdot)\|Q_\theta(l,\cdot)\big),\end{aligned} \tag{27}$$

where $C_0 = \sum_l H(P(l,\cdot))$ and $C_w = \sum_l w_l\,H(P(l,\cdot))$ are constants independent of θ. Hence minimizing $\mathcal{L}_{\mathrm{PCE}}$ is equivalent to minimizing the *weighted* KL

$$D_w(\theta) := \sum_{l=1}^{L} w_l\;\mathrm{KL}\big(P(l,\cdot)\|Q_\theta(l,\cdot)\big), \qquad \theta^\star_{\mathrm{PCE}} = \arg\min_\theta D_w(\theta). \tag{28}$$

Then, since every w_l is bounded, we can establish the relations:

$$(1-\delta)\sum_{l=1}^{L}\mathrm{KL}\big(P(l,\cdot)\|Q_\theta(l,\cdot)\big) \;\le\; D_w(\theta) \;\le\; (1+\delta)\sum_{l=1}^{L}\mathrm{KL}\big(P(l,\cdot)\|Q_\theta(l,\cdot)\big). \tag{29}$$

Dividing by L gives:

$$(1-\delta)\,D_{\mathrm{KL}}^{\mathrm{avg}}(\theta) \;\le\; \frac{D_w(\theta)}{L} \;\le\; (1+\delta)\,D_{\mathrm{KL}}^{\mathrm{avg}}(\theta). \tag{$*$}$$

Based on the fact that $\theta^\star_{\mathrm{PCE}}$ minimizes D_w, denote $\Delta_w := D_w(\theta^\star_{\mathrm{CE}}) - D_w(\theta^\star_{\mathrm{PCE}}) \ge 0$. By applying $(*)$ to both optimal parameters and subtracting, we obtain:

$$(1-\delta)\Big[D_{\mathrm{KL}}^{\mathrm{avg}}(\theta^\star_{\mathrm{CE}}) - D_{\mathrm{KL}}^{\mathrm{avg}}(\theta^\star_{\mathrm{PCE}})\Big] \;\le\; \frac{\Delta_w}{L}. \tag{30}$$

Since $\Delta_w \ge 0$ and $1-\delta > 0$; it is strictly positive whenever $\Delta_w > 0$, Therefore,

$$D_{\mathrm{KL}}^{\mathrm{avg}}(\theta^\star_{\mathrm{PCE}}) \;\le\; D_{\mathrm{KL}}^{\mathrm{avg}}(\theta^\star_{\mathrm{CE}}), \tag{31}$$

which completes the proof. □

A natural challenge emerges when applying the PCE Loss—the model tends to accurately capture the distribution of conserved regions while neglecting the distribution of variable regions. To address this issue, we demonstrate that the DIRE Loss effectively enhance the modeling in the variable regions.

Theorem 2. *For $l = 1,\dots,L$ let $P(l,a)$ denote the empirical amino-acid distribution and $Q_\theta(l,a)$ any model. When each amnio acid site is optimized independently, the minimizer is*

$$Q^\star_\alpha(l,a) = \frac{P(l,a)^{\tau_l}}{\sum_{b\in\mathcal{A}} P(l,b)^{\tau_l}}, \qquad \tau_l = \frac{\alpha w_l}{\alpha w_l + (1-\alpha)} \in (0,1). \tag{32}$$

Moreover,

$$H\big(P(l,\cdot)\big) \le H\big(Q^\star_\alpha(l,\cdot)\big) \le \log|\mathcal{A}|, \tag{33}$$

with the entropy increase largest when w_l is small (variable regions). Thus $\mathcal{L}_{DIRE}$ counter-acts the entropy suppression of $\mathcal{L}_{PCE}$ and serves as a principled regularizer on variable regions.

Proof. Since the combined loss $\mathcal{L}_\alpha$ sums over amino acid positions, we may analyze a single site independently, denoting $P(a) = P(l, a)$, $Q(a) = Q(l, a)$ and $w = w_l$. For each site we minimize, we have

$$F(Q) = \alpha w \sum_a P(a) \log \frac{P(a)}{Q(a)} + (1 - \alpha) \sum_a Q(a) \log Q(a), \tag{34}$$

subject to the normalization constraint $\sum_a Q(a) = 1$.

Introducing a Lagrange multiplier λ and setting the derivative with respect to $Q(a)$ to zero yields

$$-\frac{\alpha w P(a)}{Q(a)} \; + \; (1 - \alpha)\big(1 + \log Q(a)\big) \; + \; \lambda \; = \; 0. \tag{35}$$

Solving this equation reveals a "temperature-like" solution based on τ:

$$Q(a) \; \propto \; P(a)^\tau, \qquad \tau \; = \; \frac{\alpha w}{\alpha w + (1 - \alpha)} \in (0, 1), \tag{36}$$

which is exactly the optima $Q^\star_\alpha(l, \cdot)$ mentioned earlier.

Since $0 < \tau < 1$, this transformation always increases entropy unless P is already uniform:

$$H\big(P(l, \cdot)\big) \; \leq \; H\big(Q^\star_\alpha(l, \cdot)\big) \; \leq \; \log |\mathcal{A}|. \tag{37}$$

The entropy gain is larger when w is small (in the variable regions). Consequently, the $(1 - \alpha)$, $\mathcal{L}_{\text{DIRE}}$ term counteracts the over-confidence induced by $\mathcal{L}_{\text{PCE}}$ in variable regions, serving as an adaptive entropy-based regularizer. □

B Training and Sampling Details

Training Details We trained our model based on a Transformer T5 architecture, incorporating axial attention and task-specific modifications to enhance performance. The model consists of 12 encoder layers and 12 decoder layers, with a hidden size of 1024, 12 attention heads, and a feedforward dimension of 2048. The feedforward projection employs a gated-GELU activation function. During training, we employed the AdamW optimizer with a learning rate of 5e-5, a weight decay of 1e-5, and a polynomial decay scheduler with a 1% warmup ratio. Training was conducted on four NVIDIA A40 GPUs for up to 200,000 steps, with a batch size of 4 per device for both training and evaluation.

Sampling details The sampling process was configured with the following parameters: we generate 16 MSAs for 4 trials per generation. The sampling used a repetition penalty of 1.0, a temperature of 1.0, and top-p sampling with a threshold of 0.95. Beam search was performed with 4 beams and 1 beam group. Sampling was executed on an A40 GPU.

C Related Works

Protein Structure Prediction Protein structure prediction methods fall into three main categories: physics-based, homology-based, and deep learning approaches. Physics-based methods, such as AMBER and CHARMM, use molecular physics and energy optimization to simulate protein folding [9, 27]. While offering detailed folding insights, they are computationally expensive and sensitive to initial conditions, often yielding suboptimal results [28, 29, 30]. Homology modeling tools, like Rosetta and HHpred, use MSAs and evolutionary data to predict structures by refining templates from known experimental structures [8, 10]. These methods perform well with suitable templates but struggle with orphan proteins and low-homology families [7, 1]. Deep learning-based methods, such as AlphaFold2 and OmegaFold, use advanced neural architectures and protein templates to achieve near-experimental accuracy with greater speed and scalability [3, 5, 31]. Despite their success, they still depend on high-quality MSAs and struggle with low-homology proteins.

AlphaFold-based Enhancement Building on AlphaFold's success, researchers have developed methods to refine specific modules, aiming to improve accuracy or efficiency. These advancements can be grouped into three main categories. The first category focuses on homology expansion techniques, such as MMSeq2 and DeepMSA2, which expand the evolutionary search space to

enhance prediction accuracy. However, these methods often slow down inference despite their modest performance gains [32, 24, 33, 34]. The second category targets search acceleration, with methods like ColabFold and ESMFold bypassing the MSA search process to enhance computational efficiency. However, this speedup often results in incomplete evolutionary data, potentially reducing prediction accuracy [16, 35]. The third category leverages generative models to capture protein homology and augment input data, especially for orphan proteins and low-homology families. While promising in specific scenarios, these models struggle with extremely limited evolutionary signals, and their artificial sequences often deviate from traditional MSA distributions, limiting broader applicability [15, 13, 12, 14].

D Sequence Quality Assessments

To evaluate generated MSA quality, we designed a set of metrics focusing on MSA fidelity and diversity. These metrics aim to investigate which factors correlate most strongly with protein folding performance, especially given the lack of standardized criteria for such evaluations. Our study provides an initial exploration into this problem. Figure 3 illustrates the evaluation results.

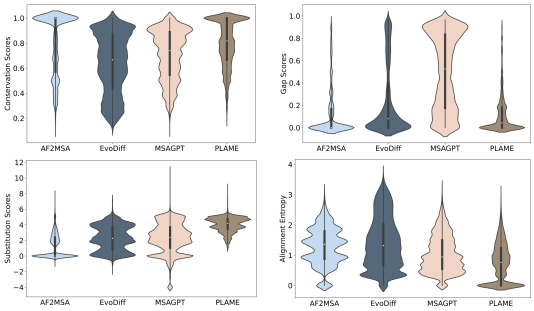


Figure 3: Comparison of sequence-based metrics for AF2 MSAs and MSAs generated by EvoDiff, MSAGPT, and PLAME.

PLAME-generated MSAs outperform other methods on fidelity metrics, with distributions closest to AF2 MSAs in Conservation Score, Gap Proportion, and Substitution Compatibility. Specifically, PLAME achieves higher Conservation Scores and Substitution Compatibility, reflecting its ability to better capture evolutionary and functional constraints. Moreover, its lower Gap Proportion indicates higher alignment completeness, which can be attributed to the latent evolutionary space providing richer homology information. These results highlight PLAME's superior alignment with the evolutionary constraints of the target protein.

We analyzed Alignment Entropy as a measure of diversity. While greater diversity is generally expected to enhance homologous information, our results show that PLAME's diversity levels are closer to those of AF2 MSAs, rather than exceeding methods like EvoDiff. This supports our assumption in 2.4 that excessive diversity can introduce noise and diminish the information enrichment effect of generated MSAs. Therefore, achieving a balance between fidelity and moderate diversity is crucial for MSA generation models that struggle to capture high-quality distributions.

E Extensive Benchmark Results

E.1 Comparison to AI-based MSA Retrieval Methods

We further compare the folding enhancement by MSAs retrieved by AI-based approaches [19]. Without filtering, DHR's performance is generally higher than AF2 MSA, EvoDiff, and MSAGPT, demonstrating the effectiveness of applying evolutionary embeddings like ESM2.

	Zero/Few	pLDDT	GDT	TMscore	RMSD	LDDT
DHR	**Zero**	63.64	0.51	0.52	12.04	0.55
DHR	**Few**	62.60	0.52	0.57	11.92	0.59
PLAME	**Zero**	66.54	0.53	0.53	11.48	0.57
PLAME	**Few**	66.08	0.54	0.58	12.14	0.60

Table 3: Comparison of folding enhancement to AI-based MSA retrieval method.

E.2 Comparison on Inference Speed and Memory Usage

To further demonstrate PLAME's efficiency, we calculated the inference time and memory cost of each method. We used ENZYME 1.2.1.50 (EC Number) with length 488 as the test case. The results show that PLAME achieved the fastest speed among all AI-based methods while consuming only 4.5GB of memory. The processing speed is comparable to traditional methods like MMSeq2 and AI-based retrieval methods like DHR. Compared to retrieval-based methods, PLAME does not require downloading or building databases in advance, nor does it need preprocessing steps. This makes it more lightweight and efficient for deployment.

Method	Time per MSA (s)	GPU Memory (Gb)
PLAME	0.10	4.5
DHR	0.16 + 358.61 (Alignment)	1.9
MMSeq2	0.48	0.0
MSAGPT	62.46	41.6
EvoDiff	478.24	4.0

Table 4: Comparison on inference speed and memory.

E.3 Ablation on HiFiAD Filtering

To demonstrate the effectiveness of HiFiAD, we conducted ablation experiments on filtering methods for other baselines. Among the benchmarks, PLAME maintains SOTA performance, and DHR-HiFiAD and PLAME show comparable performance across multiple metrics. Several metrics of EvoDiff, MSAGPT, and DHR (AI-based MSA retrieval methods) all demonstrate improvement after HiFiAD filtering compared to their original filtering methods, which directly proves the benefit of HiFiAD rules for MSA in AF-series structure prediction. This insight can help future work develop higher-quality MSA sequences.

	Zero/Few	pLDDT	GDT	TMscore	RMSD	LDDT
DHR-HiFiAD	**Zero**	66.01	0.53	0.53	11.48	0.57
DHR-HiFiAD	**Few**	66.08	0.55	0.60	12.14	0.60
EvoDiff-HifiAD	**Zero**	58.24	0.46	0.46	13.74	0.51
EvoDiff-HifiAD	**Few**	60.89	0.49	0.54	12.39	0.56
MSAGPT-HifiAD	**Zero**	60.16	0.48	0.48	12.54	0.53
MSAGPT-HifiAD	**Few**	62.63	0.52	0.57	12.18	0.59

Table 5: Ablation on HiFiAD filtering.

E.4 Ablation on Protein Length

We listed the performance of PLAME and AF2 MSA in different length ranges, and found that PLAME shows overall improvement, with the largest improvement in the 100-300 range, followed by the $>$300 and $<$100 ranges. We believe this is due to our MSA training data being mainly concentrated in the 100-300 length range.

	Length Range	pLDDT	GDT	TMscore	RMSD	LDDT
AF2 MSA	**$<$100**	71.03	0.64	0.52	7.77	0.61
AF2 MSA	**100-300**	59.50	0.49	0.53	12.46	0.54
AF2 MSA	**$>$300**	56.29	0.43	0.51	15.67	0.53
PLAME	**$<$100**	74.12	0.63	0.52	7.49	0.61
PLAME	**100-300**	65.55	0.53	0.58	11.58	0.58
PLAME	**$>$300**	58.31	0.45	0.53	16.16	0.54

Table 6: Ablation on protein length.

F Extensive Case Studies

F.1 Case Study on Successful Designs

To further explore the key pattern of the MSA augmentation, we provide a series of sequence and structure visualization in Appendix I. We select representative cases collected from different datasets and range from different lengths to comprehensive evaluate the samples.

Among these cases, we can generally observe that most generated MSA sequences maintain high similarity with the query sequence. Furthermore, the generated MSAs provide good enhancement at the originally conserved sites. This indicates that protein language models can still retain some evolutionary information even for proteins with low homology, although the diversity they can provide is more limited due to homology constraints.

Additionally, we identified several patterns in the sampled MSAs that clearly deviate from the original distribution, such as consecutive gaps (in 8ehb_F), repeated HHHHHH sequences (in 8okw_B), and repeated SSSSSSSSS (in 7xrl_A). We believe these erroneous generations are related to the autoregressive generation method, where the model tends to produce excessive hallucinations after getting trapped in incorrect local probability distributions. We also observed that these failure patterns occur more frequently in longer sequences, possibly due to insufficient training on cases with greater length. These represent an area requiring further improvement.

F.2 AlphaFold3 Folding Experiment

To further validate PLAME is an universal MSA generation framework for different folding software, we conduct extensive experiment on AlphaFold3 [5]. We use the same generated MSAs filtered by HiFiAD, augmenting the original MSAs for folding enhancement. Results follow the same trend as

	pLDDT	GDT	TMscore	RMSD	LDDT	pTM
AF2 MSA	68.872	0.578	0.596	10.818	0.617	0.529
PLAME	70.887	0.578	0.595	10.740	0.623	0.539

Table 7: Comparison of folding enhancement based on AlphaFold3

AF2 Mode 3. The stronger the folding baseline, the smaller the performance gain. Improvements are mainly in pLDDT and RMSD, indicating our enhanced MSA primarily improves local structural regions while preserving the global protein architecture.

F.3 Folding Enhancement on Average Proteins

To probe the effectiveness of PLAME on average proteins, we firstly build a dataset from PDB validation set with 36 proteins. These protein MSAs don't have sequence similarity over 90% compared to the PLAME training set. We randomly employ 32 MSAs for each protein and augment them with designed MSAs after HiFiAD filtering. The results are shown in Table 8. From the

	pLDDT	GDT	TMscore	RMSD	LDDT	pTM
AF2 MSA	83.156	0.767	0.785	5.243	0.753	0.718
PLAME	83.328	0.775	0.795	5.028	0.757	0.723

Table 8: Comparison of folding enhancement on average proteins

experimental results, the effects of augmentation align with our initial assumptions, demonstrating modest improvements. While the overall topological structure remains unchanged, minor adjustments can be observed in the structural details. As reported in MSAGPT, performance gains approach saturation between 16 and 32 augmentations. The relatively small improvements observed when applying our method to the average protein MSA can be attributed to the fact that these original MSAs already provide sufficient evolutionary information to AlphaFold2's MSA Transformer, thus limiting the potential impact of additional augmentation.

F.4 Further Ablation on MSA Filtering

We further validate the effectiveness of filtered high-quality MSAs by comparing the performance with the more randomly selected MSAs (64 for each protein). From Table 9 and 2, We can observe a

	pLDDT	GDT	TMscore	RMSD	LDDT	pTM
More Random MSAs	63.620	0.512	0.533	12.692	0.563	0.473
HiFiAD	66.349	0.534	0.553	11.755	0.581	0.506

Table 9: Comparison of folding enhancement based on different filterings.

slight performance enhancement compared to Random-16 filtering approach according to pLDDT and LDDT. Conversely, the performance on global metric decreases. From the results, more co-evolutionary information may lead to better local geometric conformation, but it will disturb the modeling of the global conformations due to the bias during generation.

F.5 Failure Case Analysis

Other than analyzing successful cases, we analyzed four representative failure cases (3bog_B, 7sxb_A, 8gzu_AN, 8gzu_T3) with the largest performance drops, which includes three zero-shot and one few-shot examples. From the detailed results, we observe a clear mismatch between global metric, including GDT, TMScore, and RMSD, and local metric, including pLDDT, LDDT, and pTM on 3bog_B and 8gzu_T3. It is consistent with the metric discrepancies we observed in the main experiment. Among the visualized MSA cases, we observed that generated MSAs contained

	pLDDT	GDT	TMscore	RMSD	LDDT	pTM
AF2 MSA						
3bog_B	41.493	0.150	0.130	22.443	0.148	0.129
7sxb_A	84.931	0.739	0.757	2.559	0.661	0.753
8gzu_AN	58.189	0.390	0.488	17.630	0.700	0.406
8gzu_T3	59.533	0.591	0.668	14.030	0.659	0.597
PLAME						
3bog_B	32.918	0.169	0.148	17.522	0.158	0.118
7sxb_A	53.956	0.358	0.358	9.988	0.369	0.359
8gzu_AN	51.542	0.393	0.491	17.238	0.513	0.414
8gzu_T3	55.169	0.377	0.480	20.930	0.691	0.394

Table 10: Comparison of folding enhancement on failure cases.

extremely similar sequences (>90% similarity). Specifically, these high-similarity sequences caused all sites to appear more conserved, resulting in a lack of covariation patterns necessary for AlphaFold2 to infer structural contacts. This pattern was evident across all four cases. Notably, for 3bog_B and 8gzu_T3, the generated high-similarity MSAs further enhanced the conservation of already conserved regions, which consequently led to improvements in global metrics.

F.6 De novo Protein Folding Enhancement

We conduct further experiments on De Novo protein cases, where almost of them are orphan. Examples of de novo proteins include 8SK7 (RFDiffusion), 8TNM/8TNO (Chroma), and 8CYK (ProteinMPNN). We followed the same augmentation pattern as the main experiment. From Table

	pLDDT	GDT	TMscore	RMSD	LDDT	pTM
AF2 MSA	89.27	0.886	0.904	1.658	0.781	0.800
HiFiAD	88.33	0.924	0.940	1.483	0.824	0.800

Table 11: Comparison of folding enhancement on de novo proteins.

11, we observed that PLAME experiences a slight decrease in pLDDT scores while simultaneously showing improvements in other metrics. The generated MSA visualizations in Figures 4 and 5 reveal that most generated sequences maintain $> 70\%$ similarity to the query sequences. This phenomenon

may be attributed to these test cases being highly Out-Of-Distribution (OOD) relative to the training dataset. Nevertheless, the diverse sampling strategy still effectively enhances the profile information of orphan proteins, resulting in substantial performance improvements. Furthermore, we visualized specific local regions where PLAME achieves superior alignment performance as measured by TMscore. Analysis revealed that across all augmented profiles, these high-performing local regions exhibit remarkable conservation, suggesting a strong correlation between sequence conservation patterns and structural alignment quality.

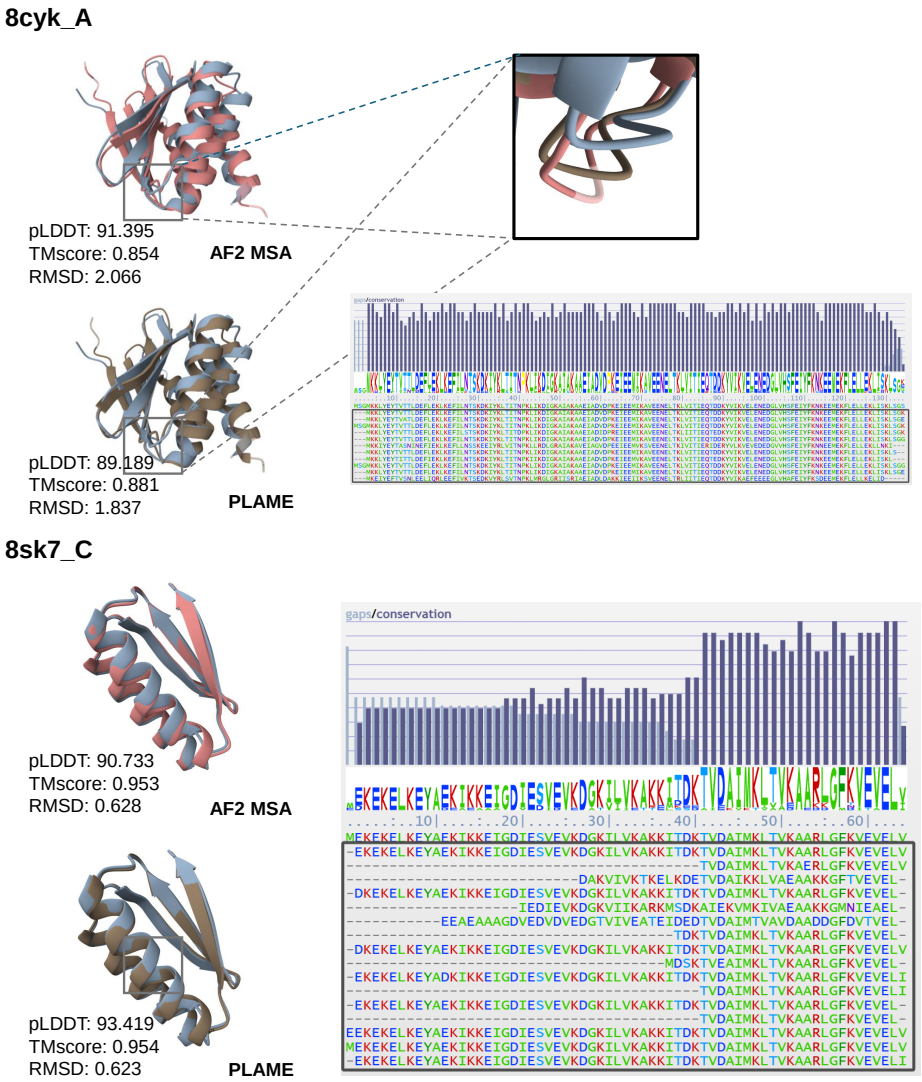


Figure 4: Comparison of structure enhancement of De Novo proteins.

G Discussion

G.1 Limitations

Recent advancements in MSA generation models have shown promising results in enhancing protein folding predictions. However, several challenges remain to be addressed for broader applications and improved performance. **1) Limited quality** by current model architectures, data constraints, and generation strategies, such as relying on small MSA prompts, hinders the overall richness and informativeness of the generated MSAs. Future methods should focus on constructing more

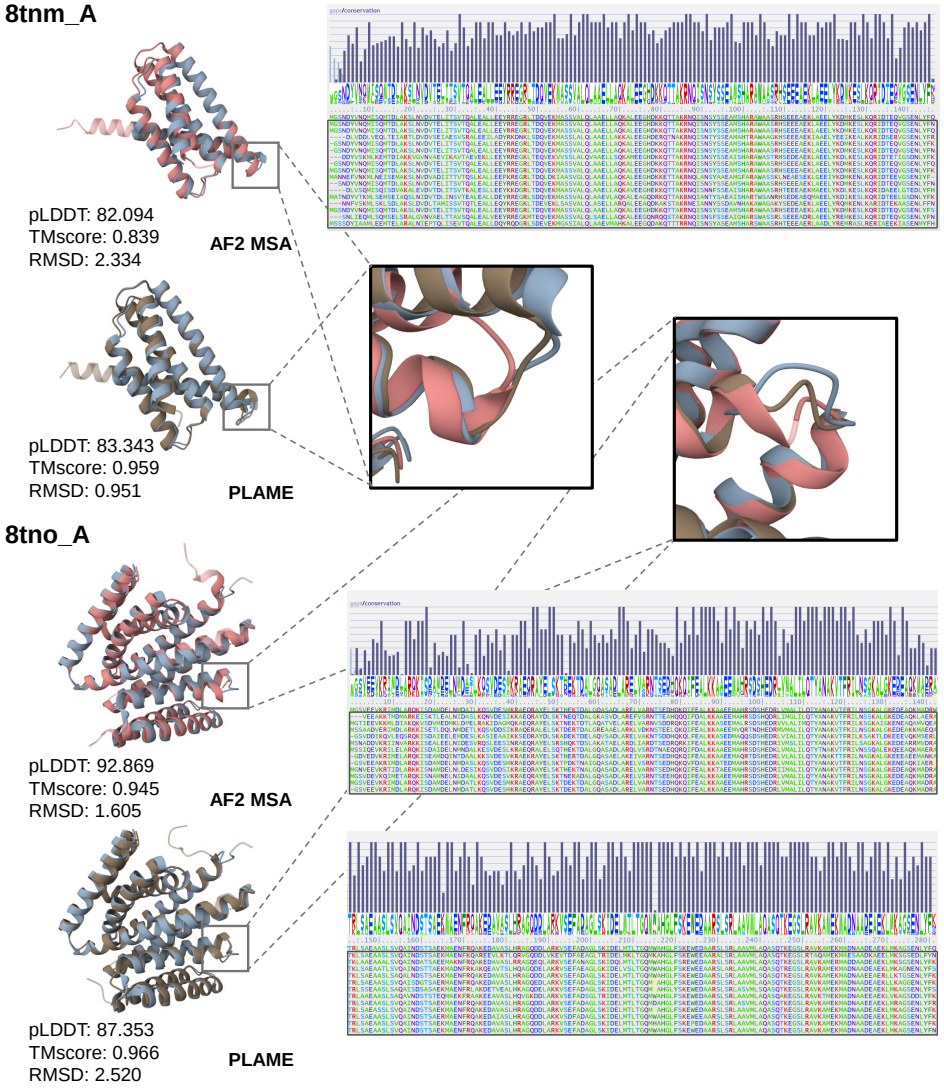


Figure 5: Comparison of structure enhancement of De Novo proteins.

expressive evolutionary latent spaces to better capture the complexity of protein sequence relationships and improve the informativeness of generated MSAs. **2) Distribution gaps** still exist between the diversity and quality of generated MSAs and their natural counterparts, limiting their utility in broader applications. While current methods show potential in folding tasks, future models should focus on zero-shot generation capabilities to produce MSAs with distributions closer to natural MSAs, enabling broader applications such as conserved residue identification, mutation effect prediction, and functional annotation. **3) Assessing MSA quality** remains an unresolved issue, as current evaluations primarily rely on downstream folding performance to infer quality. Developing direct and robust quality assessment metrics will be crucial for systematically evaluating and improving MSA generation methods, enabling the selection of high-quality MSAs for specific applications and paving the way for next-generation models with enhanced accuracy, broader applicability, and greater biological relevance.

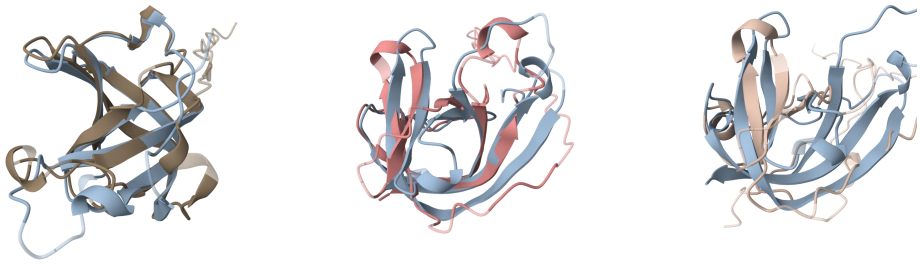
G.2 Social Impact

PLAME advances multiple sequence alignment (MSA) generation for proteins, offering significant potential benefits for protein structure prediction and engineering applications. This capability

could accelerate therapeutic protein development, enhance drug design processes, and facilitate the exploration of novel functional protein sequences. However, we acknowledge important limitations and risks associated with this technology. The model may occasionally generate hallucinated sequences that could mislead downstream protein design efforts if not properly validated. Furthermore, like many powerful biotechnology tools, there exists potential for misuse in designing harmful biological entities.

H Structure Comparison Visualization

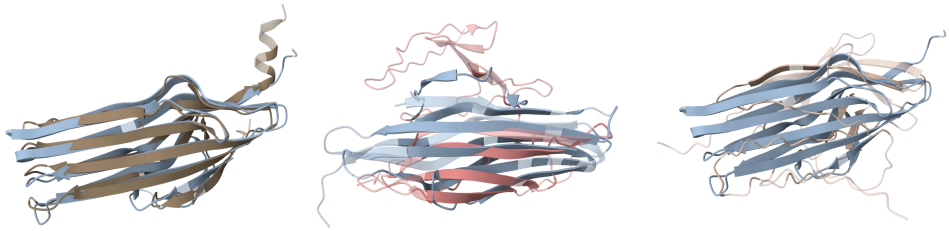
pdb_id: 8ehb_F



PLAME	MSAGPT	AF2MSA
pLDDT: 75.64 TMscore: 0.749 RMSD: 3.218	pLDDT: 41.35 TMscore: 0.563 RMSD: 4.462	pLDDT: 36.25 TMscore: 0.359 RMSD: 9.653

Figure 6: Structure comparison visualization of 8ehb_F.

pdb_id: 8okh_B



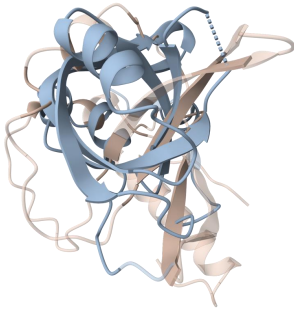
PLAME	MSAGPT	AF2MSA
pLDDT: 69.67 TMscore: 0.812 RMSD: 2.774	pLDDT: 32.02 TMscore: 0.205 RMSD: 19.49	pLDDT: 28.67 TMscore: 0.198 RMSD: 21.09

Figure 7: Structure comparison visualization of 8okh_B.

pdb_id: 8b4k_C

PLAME

pLDDT: 79.43
TMscore: 0.743
RMSD: 6.10

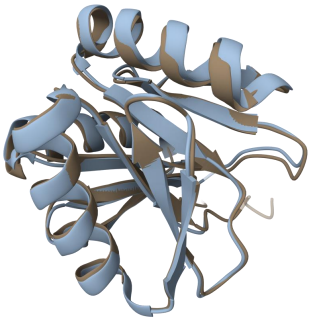


AF2MSA

pLDDT: 31.49
TMscore: 0.209
RMSD: 15.37

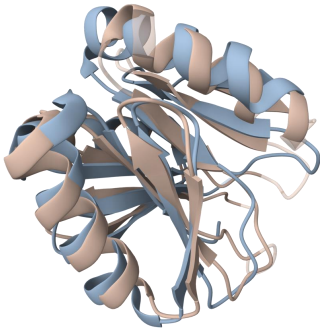
Figure 8: Structure comparison visualization of 8b4k_C.

pdb_id: 8fjf_A



PLAME

pLDDT: 91.56
TMscore: 0.974
RMSD: 0.783



AF2MSA

pLDDT: 64.56
TMscore: 0.734
RMSD: 3.193

Figure 9: Structure comparison visualization of 8fjf_A.

pdb_id: 8eoz_B

PLAME

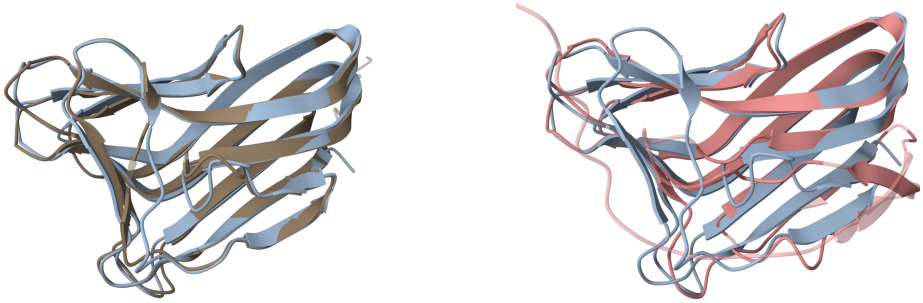
pLDDT: 88.24
TMscore: 0.958
RMSD: 0.127

MSAGPT

pLDDT: 48.10
TMscore: 0.290
RMSD: 15.338

Figure 10: Structure comparison visualization of 8eoz_B.

pdb_id: 8okw_B



PLAME

pLDDT: 86.18
TMscore: 0.945
RMSD: 1.408

MSAGPT

pLDDT: 52.90
TMscore: 0.658
RMSD: 12.459

Figure 11: Structure comparison visualization of 8okw_B.

I Augmented MSA Visualization

To provide an intuitive understanding of the MSAs generated by PLAME, we selected several representative cases for visualization. These cases demonstrate consistent improvements in folding accuracy compared to the MSAs provided by AF2 and cover a range of sequence lengths, including short (<100), medium (100-300), and long (>300) sequences, as well as cases under few-shot and zero-shot settings. For each visualization, the generated MSAs are highlighted with a black box. Additionally, the upper portion of each figure presents conservation information alongside the corresponding gap information. The protein information is provided in the left-top corner at each figure.

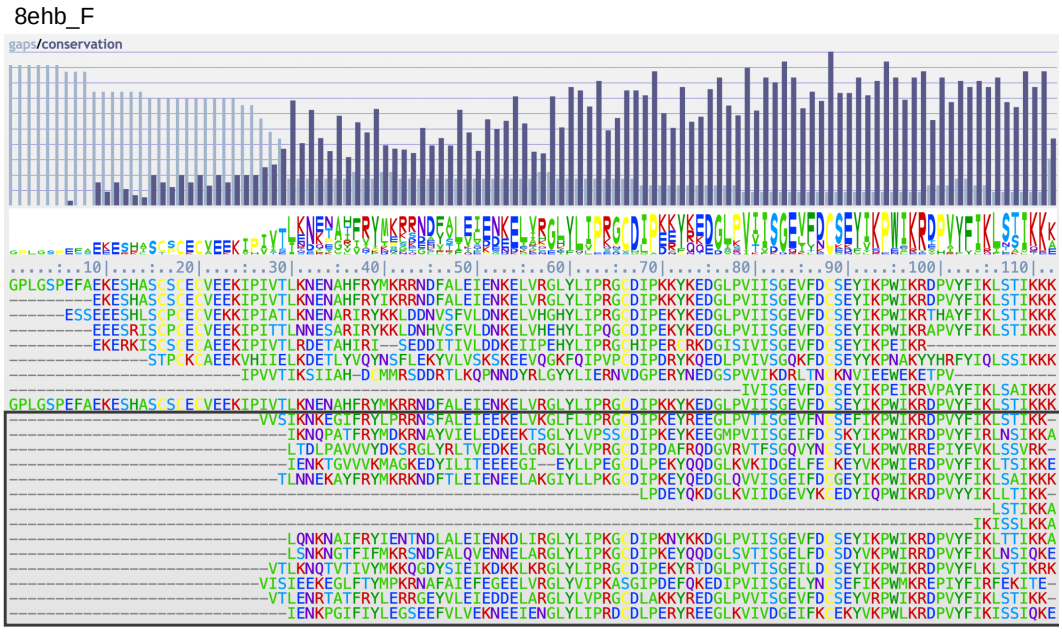


Figure 12: Augmented MSA visualization of 8ehb_F.

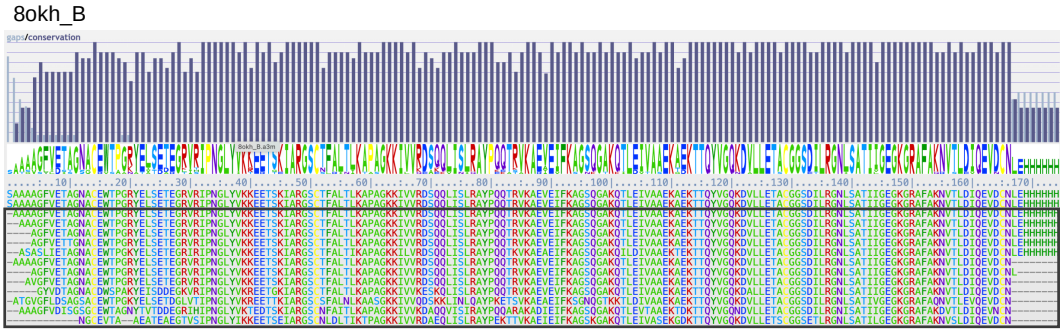


Figure 13: Augmented MSA visualization of 8okh_B.

8okw_B

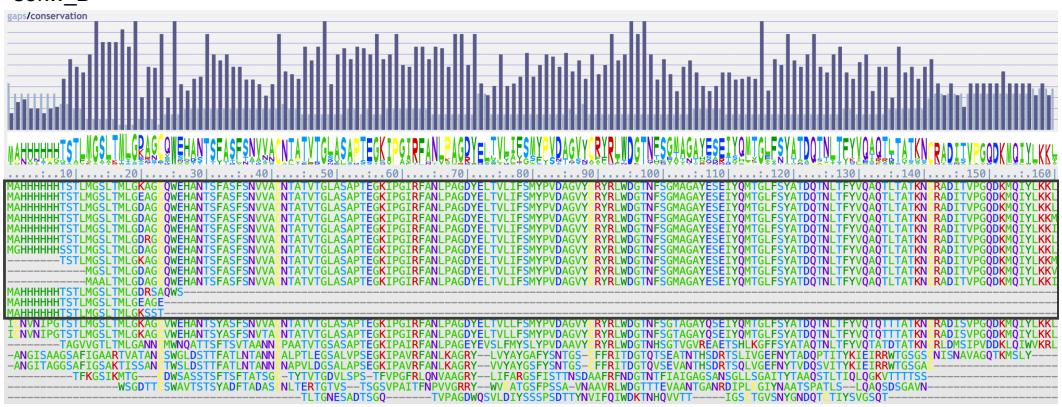


Figure 14: Augmented MSA visualization of 8okw_B.

8fih_C

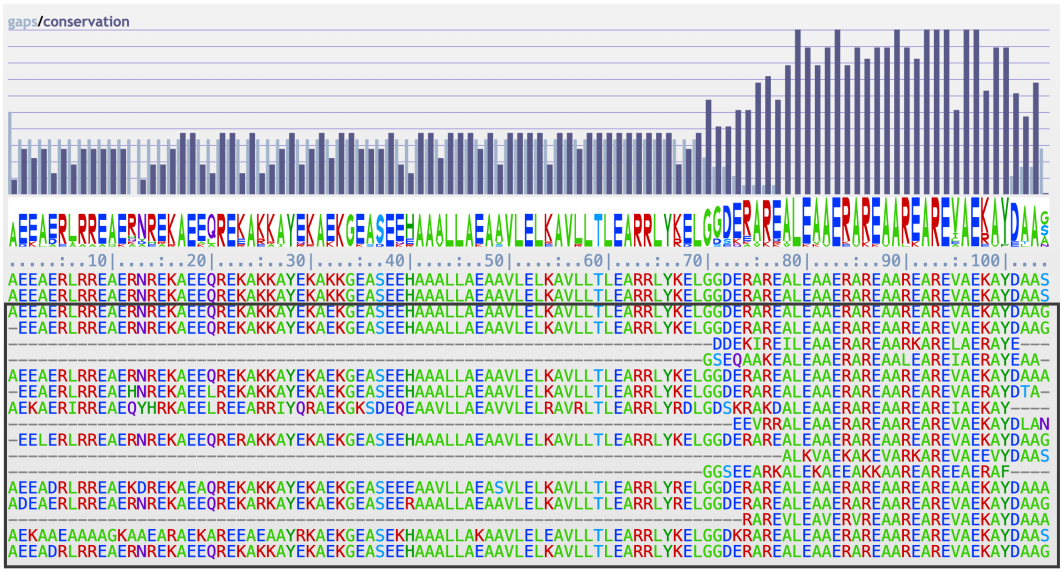


Figure 15: Augmented MSA visualization of 8fih_C.

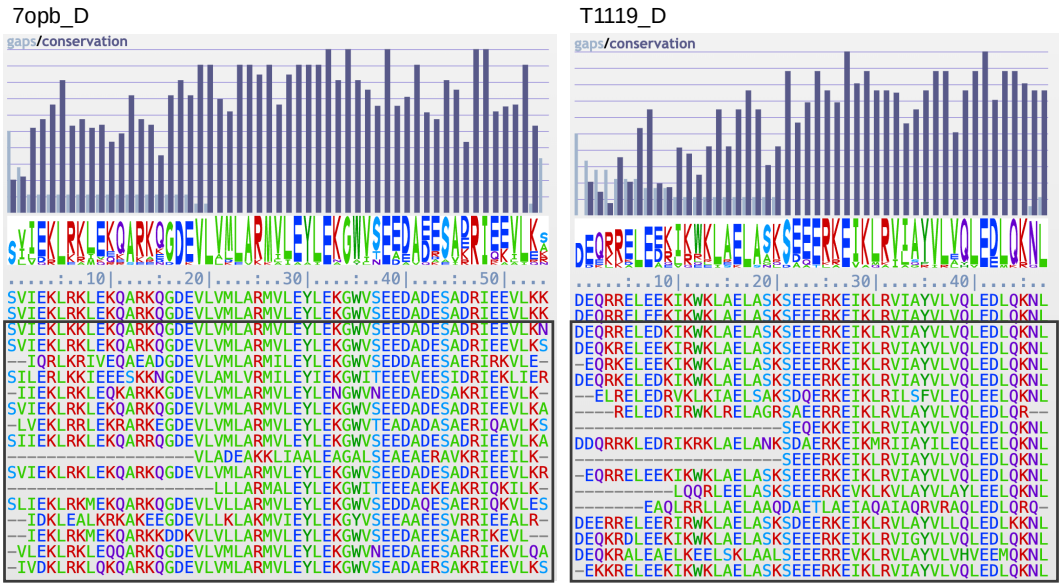


Figure 16: Augmented MSA visualization of 7opb_D and T1119_D.

Figure 17: Augmented MSA visualization of 7xr1_A.

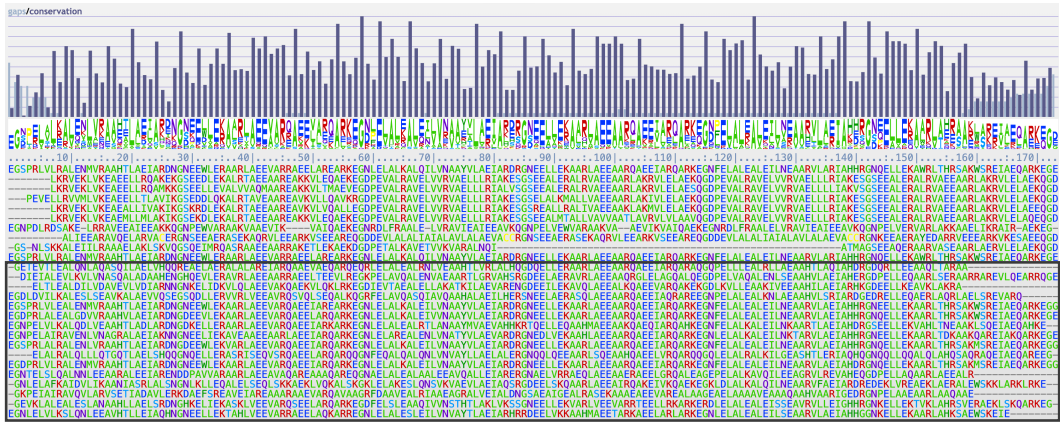


Figure 18: Augmented MSA visualization of 8e0n_F.

J Failure Case MSA Visualization

7sxb_A

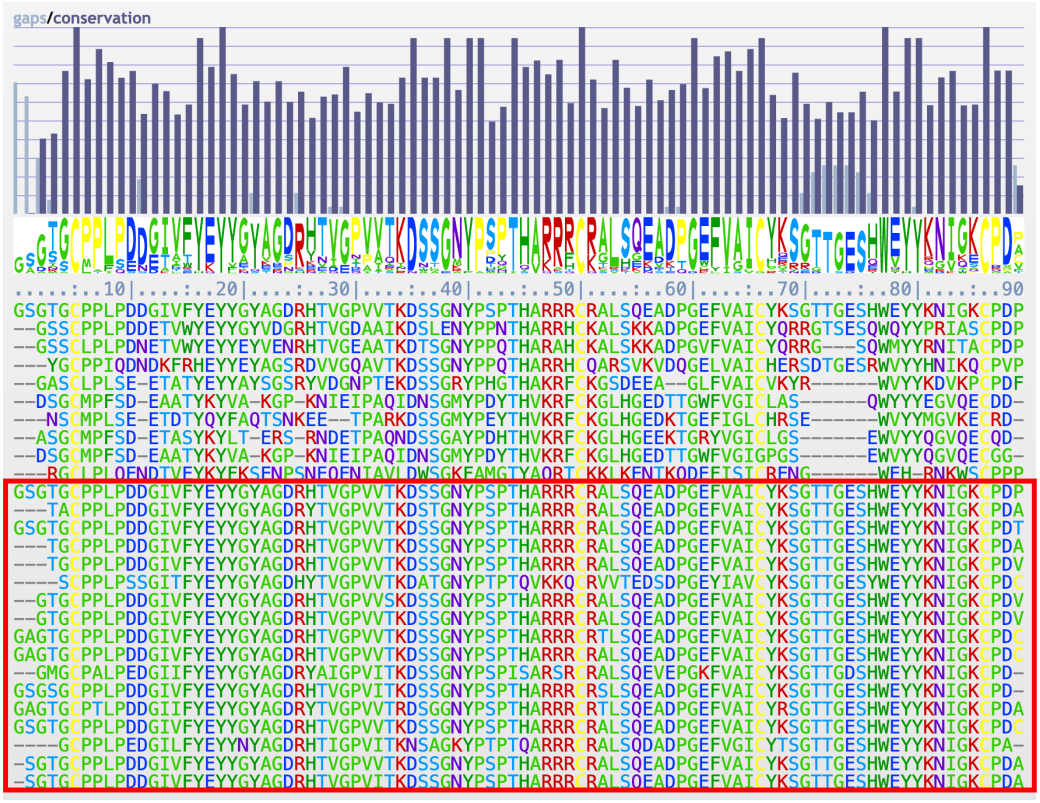


Figure 19: Failure Case MSA visualization of 7sxb_A.

8gzu_T3

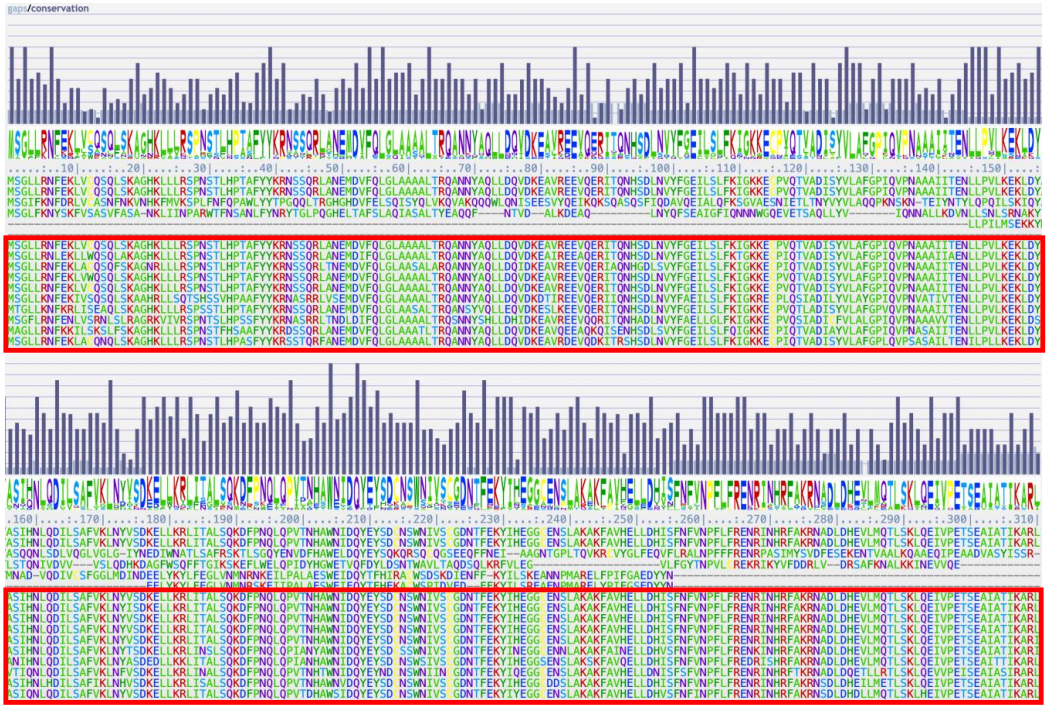


Figure 20: Failure Case MSA visualization of 8gzu_T3.

3bog_B

Figure 21: Failure Case MSA visualization of 3bog_B.

8gzu_AN

Figure 22: Failure Case MSA visualization of 8gzu_AN.