MSAFlow: a Unified Approach for MSA Representation, Augmentation, and Family-based Protein Design

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multiple Sequence Alignments (MSAs) provide fundamental information about protein evolutionary trajectories and play crucial roles in downstream tasks such as augmentation and family-based design . However, constructing high-quality MSAs requires significant computational resources to query natural protein databases, and traditional techniques fail to provide relevant data for proteins with limited evolutionary information. While deep learning approaches have shown promise in MSA construction and augmentation, they fail to capture rich distributional information while preserving permutation invariance. MSAFlow addresses these limitations using a Statistical Flow Matching model conditioned on compressed latent MSA representations to generate sequences that would likely belong to the target MSA. This approach captures distributional information while augmenting shallow MSAs and maintaining permutation invariance. Experiments confirm that MSAFlow generates MSAs with performance comparable to traditional methods on family-based design tasks. The model outperforms existing machine learning augmentation tools while achieving very low inference time and memory efficiency despite being lightweight and trained on smaller datasets. MSAFlow enables family-based protein design for enzymes and synthetic MSA generation through latent diffusion. Extensive ablation studies validate the effectiveness of model design components. Overall, MSAFlow provides a robust and efficient framework for MSA representation and integration in downstream applications.

1 Introduction

Multiple Sequence Alignments (MSAs) provide fundamental information about protein evolutionary trajectories and play crucial roles in downstream tasks such as augmentation and family-based design. MSAs represent collections of homologous proteins that delineate the evolutionary history of a single query sequence, enabling models to identify conserved regions and detect evolutionary couplings. Moreover, MSAs carry significant information about functional sites within the query sequence; for instance, comparing sequences across a family of enzymes can reveal conserved active site residues. However, constructing high-quality MSAs requires significant computational resources to query natural protein databases. While traditional statistical search methods such as HHBlits [1], MMSeqs [2], and JackHMMER [3] can accurately identify evolutionarily-related sequences, they incur significant computational costs and traditional techniques fail to provide relevant data for proteins with limited evolutionary information.

This challenge has been partially addressed by Dense Homology Retriever (DHR) [4], which leverages pretrained embeddings from protein language models to identify homologous sequences more efficiently and with greater sensitivity. Several other models, including MSAGenerator [5], MSAGPT

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

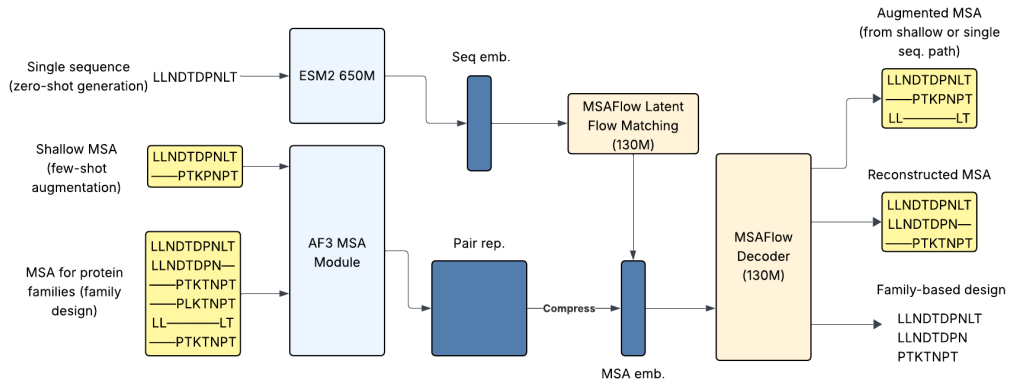

Figure 1: **General framework of MSAFlow.** Our approach supports three complementary pathways: (1) zero-shot generation from a single sequence using ESM2 embeddings, (2) few-shot augmentation of shallow MSAs, and (3) family-based design given MSAs embedded through the AF3 MSA Module and reconstructed through MSAFlow Decoder. All pathways leverage the latent flow-matching and decoder architecture to generate augmented or compressed MSAs, enabling both the enhancement of limited evolutionary information and the efficient representation of deep alignments.

[6], and EvoDiff [7], have subsequently emerged, employing autoregressive or discrete diffusion frameworks to model the joint distribution of multiple sequences in MSAs. While deep learning approaches have shown promise in MSA construction and augmentation, they fail to capture rich distributional information while preserving permutation invariance.

However, these methods typically utilize 2D positional encodings to represent row-wise and column-wise information present in MSAs. These approaches fail in critical aspects: they are substantially memory-intensive due to the $O(N^2)$ space complexity of self-attention operations, further exacerbated by the 2D nature of MSAs, and they lack permutation invariance, naively prioritizing certain sequences without employing permutation-invariant aggregation techniques. Furthermore, current MSA generation models rely solely on existing MSA sequence information for generation, limiting their effectiveness in shallow MSA enhancement.

To address these limitations, we introduce MSAFlow, which addresses these limitations using a Statistical Flow Matching [8] (SFM) model conditioned on compressed latent MSA representations to generate sequences that would likely belong to the target MSA. The MSAFlow framework first employs the AlphaFold3 [9] (AF3) MSAModule to generate a latent MSA embedding, which aggregates protein MSA information into its corresponding pair representation. We further compress this representation into a single-sequence representation through mean pooling across the second dimension. This embedding subsequently serves as conditional information for the Statistical Flow Matching model trained to reconstruct sequences from the original MSA. This approach captures distributional information while augmenting shallow MSAs and maintaining permutation invariance during reconstruction while enabling latent flow matching on the MSA embedding itself.

Experimental results demonstrate that MSAFlow generates MSAs with performance comparable to traditional methods on family-based design tasks. The model outperforms existing machine learning augmentation tools while achieving very low inference time and memory efficiency despite being lightweight and trained on smaller datasets. We evaluate MSAFlow on its ability to reconstruct MSAs from compressed latent representations, testing both the expressivity of the embeddings and the model's capacity to interpolate across the entire evolutionary space of proteins. Additionally, we utilize MSAFlow to augment existing shallow MSAs and generate synthetic MSAs for single sequences with limited evolutionary data to enhance structure prediction. MSAFlow enables family-based protein design for enzymes and synthetic MSA generation through latent diffusion, providing a robust and efficient framework for MSA representation and integration in downstream applications.

2 Preliminaries

Multiple Sequence Alignments (MSAs) are mathematically represented as $\mathcal{S} = \{s_1, s_2, ..., s_M\}$ where each sequence $s_i \in \mathcal{A}^L$ consists of amino acids and gaps from alphabet $\mathcal{A}$, aligned to a reference sequence s_{ref} of length L. Despite containing hundreds to thousands of sequences, we hypothesize that the functional and evolutionary information within an MSA can be **compressed into a continuous latent representation** that captures the essential characteristics of the protein family.

This compression necessitates a permutation-invariant encoding method to avoid bias from sequence ordering. Formally, we seek an encoder $h_\phi : \mathcal{S} \to \mathbb{R}^d$ such that $h_\phi(\mathcal{S}) = h_\phi(\pi(\mathcal{S}))$ for any permutation π of the sequences in $\mathcal{S}$. We leverage the AlphaFold3 (AF3) MSAModule architecture, which provides a computationally efficient framework for embedding evolutionary information [9]. The AF3 MSAModule processes an MSA by computing a position-wise outer product for each sequence s_i with the reference sequence, resulting in pairwise representations $P_i \in \mathbb{R}^{L \times L \times h_{\text{pair}}}$. These representations are averaged across all sequences:

$$P_{\text{avg}} = \frac{1}{M} \sum_{i=1}^{M} P_i \tag{1}$$

The averaged representation is then processed through multiple triangle self-attention blocks to produce a refined pair representation $P_{\text{refined}} \in \mathbb{R}^{L \times L \times H}$. We utilize Protenix [10], a pretrained variant of AF3, to generate these embeddings for MSAs from the OpenFold dataset [11]. The resulting pair representation serves as our compressed MSA embedding $m = h_\phi(\mathcal{S}) \in \mathbb{R}^{L \times L \times H}$.

3 Method

3.1 Flow-matching based autoencoder for MSA representation

3.1.1 Probabilistic Framework

We view our model as a conditional generator over the sequence distribution of a protein family. Given an MSA $\mathcal{S}$ and its embedding $m = h_\phi(\mathcal{S})$, the decoder aims to reconstruct sequences consistent with the family. Let $\tilde{\mathcal{S}} = \{s_1, \ldots, s_n\}$ be n sequences drawn uniformly without replacement from $\mathcal{S}$. We model

$$p_\theta(\tilde{\mathcal{S}} \mid m) = \prod_{i=1}^{n} p_\theta(s_i \mid m), \tag{2}$$

which is permutation-invariant by construction. The decoder $p_\theta(s \mid m)$ represents the probability of sampling a sequence s compatible with m.

3.1.2 Statistical Flow Matching for MSA sequence decoding

To instantiate $p_\theta(s \mid m)$ for discrete (categorical) sequences, we adopt Statistical Flow Matching (SFM) [12], which learns a continuous Riemannian flow over the statistical manifold of categorical distributions equipped with Fisher-Rao metric. Concretely, each sequence in the MSA is treated as a sample of the target distribution. We operate in the probability simplex $\Delta^{|\mathcal{A}| \times L}$, where each position in the sequence is represented by a one-hot categorical distribution μ over amino acids.

Following SFM, we construct flow paths along geodesics on the positive orthant of the unit sphere by applying the mapping: $\pi : x = \pi(\mu) = \sqrt{\mu}$. SFM demonstrated that such a mapping to the unit sphere preserves the metric, which coincides with the canonical spherical geometry. Therefore, we can operate on the unit sphere with the standard spherical geometry. Mathematically, given a sequence s_i from the MSA and its corresponding categorical representation $x_1 = \pi(\mu_1)$ (e.g., one-hot encoding) and the noise representation $x_0 = \pi(\mu_0)$, the time-dependent interpolation follows:

$$x_t = \exp_{x_0}(t \cdot \log_{x_0}(x_1)) \tag{3}$$

where exp and log are the spherical exponential and logarithm maps on the manifold, respectively, and can be calculated in closed form as

$$\exp_x(u) = x \cos \|u\|_2 + \frac{u}{\|u\|_2} \sin \|u\|_2, \tag{4}$$

$$\log_x(y) = \frac{\arccos(\langle x, y \rangle)}{\sqrt{1 - \langle x, y \rangle^2}} (y - x - \langle x, y - x \rangle x), \tag{5}$$

After transforming back to the simplex with $\mu_t = \pi^{-1}(x_t)$, the interpolation in Equation 3 traces the geodesic between μ_0 and μ_1 with respect to the Fisher information metric, ensuring we follow the shortest path on the statistical manifold. The corresponding vector field for this mapped geodesic flow is given by:

$$u_t(x_t|x_0, x_1) = \frac{\log_{x_t}(x_1)}{1 - t}. \tag{6}$$

Instead of an unconditional model, our MSAFlow decoder employs a conditional parameterization where $v_\theta(x_t|m, t)$ is trained to approximate the vector field conditioning on the MSA embedding $m = h_\phi(\mathcal{S})$:

$$\mathcal{L}_{\text{SFM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1], s_i \sim \mathcal{S}, \mu_0 \sim \pi_* p_0, \mu_1 \sim \pi_* \delta(s_i)} \left[\|v_\theta(x_t|m, t) - u_t(x_t|x_0, x_1)\|^2 \right] \tag{7}$$

where π_* denotes the pushward of applying the mapping π, x_t is obtained via the geodesic interpolation, and $\delta(s_i)$ represents the categorical distribution corresponding to sequence s_i (typically a one-hot encoding) in an MSA. During sampling, we first follow the learned marginal vector field on the sphere to obtain x_1, then discrete generations of MSAs can be sampled from the categorical distribution $\mu_1 = \pi^{-1}(x_1)$.

3.1.3 Model Architecture and Implementation

We implement the vector field model v_θ using a modified conditional Diffusion Transformer (DiT) architecture. Since the output of the AF3 MSAModule is the pair representation of dimension $L \times L \times H$, we first compress it along the second dimension through mean pooling to obtain a sequence-level representation of dimension $L \times H$:

$$m_{\text{seq}} = \frac{1}{L} \sum_{j=1}^{L} m_{:,j,:} \in \mathbb{R}^{L \times H} \tag{8}$$

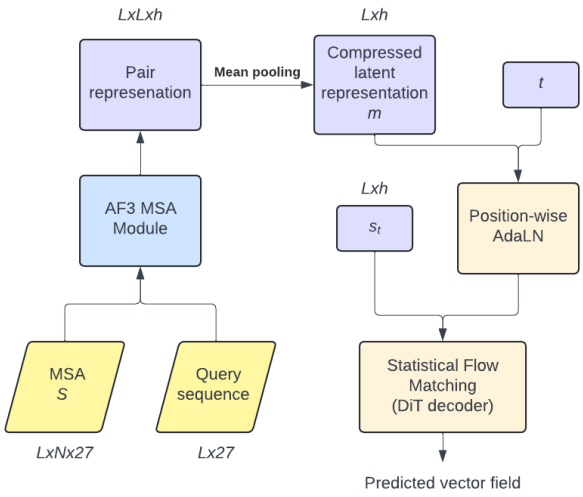

Figure 2: DiT architecture for MSAFlow decoder.

This compressed representation serves as conditional information for the DiT model, which consists of 12 transformer blocks with a hidden dimension of 768, totaling approximately 130M parameters. The architecture incorporates sinusoidal time embeddings for the diffusion timestep t, token embeddings for each amino acid position, conditional embeddings from the compressed MSA representation, and multi-headed self-attention blocks with adaptive layer normalization. Notably, the MSA embedding conditioning is applied per-residue through a position-wise AdaLN to achieve residue-level control. At inference time, we sample sequences by starting with random noise $x_1 \sim \text{Uniform}(\mathcal{A})$ and iteratively applying:

$$x_{t-\Delta t} = x_t - v_\theta(x_t|m, t) \cdot \Delta t \tag{9}$$

for timesteps $t = 1, 1 - \Delta t, 1 - 2\Delta t, ..., 0$, where Δt is a small step size (typically 0.01). At $t = 0$, we obtain the final sequence by taking the argmax over the amino acid probabilities at each position.

3.2 Conditional latent flow matching for MSA embedding generation

While our decoder model generates sequences from MSA embeddings, we also develop a complementary approach to generate synthetic MSA embeddings themselves. This enables us to create artificial MSAs for proteins with limited evolutionary data (e.g., de novo proteins and antibodies).

3.2.1 Problem Formulation

Let $z_1 = h_\phi(\mathcal{S}) \in \mathbb{R}^{L\times H}$ be the compressed MSA embedding for a reference sequence s_{ref}, and let $e = g_\psi(s_{\text{ref}}) \in \mathbb{R}^{d_e}$ be its ESM embedding. We aim to learn a conditional generative model $p_\theta(z_1|e)$ that can produce plausible MSA embeddings given only the reference sequence embedding.

3.2.2 Latent Flow Matching

We train a *conditional rectified flow* that maps a standard Gaussian $z_0 \sim \mathcal{N}(0, I)$ to the distribution of MSA embeddings $p(z \mid e)$ conditioned on the ESM embedding e. We use a straight-line path $z_t = (1-t)\, z_1 + t\, z_0$ from target z_1 (the ground-truth MSA embedding) to noise z_0, whose reference velocity is the constant field $u_t^\star(z_t; z_0, z_1) = z_0 - z_1$. A time-dependent, conditional velocity $v_\theta(z_t, e, t)$ is learned by least-squares flow matching:

$$\mathcal{L}_{\text{RFM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, z_0\sim\mathcal{N}(0,I),\, z_1} \left\| v_\theta(z_t, e, t) - (z_0 - z_1)\right\|_2^2,$$

which provides a simple, stable objective without explicit score estimation.

3.2.3 Generative Sampling Process

At inference, we draw $z_0 \sim \mathcal{N}(0, I)$ and integrate the learned conditional velocity backward from $t{=}1$ to $t{=}0$ with an explicit Euler solver. By default we use the deterministic probability-flow ODE ($T{=}0$); optionally, we add isotropic noise with temperature $T \in [0, 1]$ to trade fidelity for diversity:

$$z_{t-\Delta t} = z_t - v_\theta(z_t, e, t)\, \Delta t \;+\; T\sqrt{\Delta t}\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I).$$

Empirically, smaller T (e.g., $T{<}0.5$) improves alignment to e, while larger T increases sample diversity. Full SDE variants and discretization details follow [13] and are deferred to the Appendix.

3.2.4 End-to-End MSA Generation Pipeline

Our complete framework enables two complementary paths for MSA generation (as shown in Figure 1), each tailored to specific protein scenarios:

MSA Compression and Reconstruction: For deep MSAs with abundant evolutionary information, we first compress the multidimensional sequence information through the AF3 MSAModule into a compact latent representation. This compressed embedding effectively captures the evolutionary and functional signals present in the original MSA. We then use our SFM decoder to selectively reconstruct sequences, maintaining the key evolutionary characteristics while reducing redundancy.

Zero-shot MSA Generation: For orphan or de novo proteins with limited evolutionary data, we first generate the ESM embedding of the single available sequence. Our latent diffusion model then transforms this single-sequence representation into a synthetic MSA embedding that emulates the evolutionary diversity typically found in natural protein families. Finally, we decode multiple diverse sequences from this embedding using our SFM decoder, effectively bootstrapping evolutionary information where none previously existed.

Family-based Design: To perform family-based design for enzymes, we first gather all sequences belonging to the enzyme class of a given query. These sequences are compressed into a latent representation using the MSAModule distilled from AF3. Our SFM decoder then generates new sequences conditioned on this latent embedding, effectively producing candidates that are highly likely to belong to the original enzyme class. Because the generated sequences may include gaps, we can support both variable-length and fixed-length design: gaps can be ignored when constructing the final sequence, enabling flexible design strategies.

This approach combines both MSA compression and generation capabilities in a unified framework. For data-rich scenarios, our method enables efficient information extraction from deep MSAs while

preserving their evolutionary signals. For data-limited proteins, it allows the creation of synthetic alignments that capture potential evolutionary diversity. The integration of these complementary pathways addresses a fundamental limitation in protein analysis by extending evolutionary context to proteins that previously lacked sufficient homologous sequences, potentially improving downstream structure prediction, functional annotation tasks, and family-based design ability.

4 Experiments

4.1 Benchmarking MSA Autoencoding

We evaluate the reconstruction ability of our model on 50 proteins released by CAMEO on May 10, 2025, where the ground truth MSA is generated using the same procedure as described in [10]. We then compute the embedding for each MSA via the AF3 MSAModule, and generate 32 sequences given each latent MSA representation. We find that the relatively shallow MSAs generated by our model through this method come close to matching the deep, ground-truth MSAs in terms of pLDDT (87.8 vs. 91.6) and TM-scores (0.83 vs. 0.89) while only consuming 6.5% of the overall bits required to represent a deep MSA (this is for an average sequence length of 365 and number of alignments being more than 7,000 from the CAMEO dataset. We perform conditional generation given an embedding of 16-bit floats with an average size of 365×128 from the CAMEO dataset).

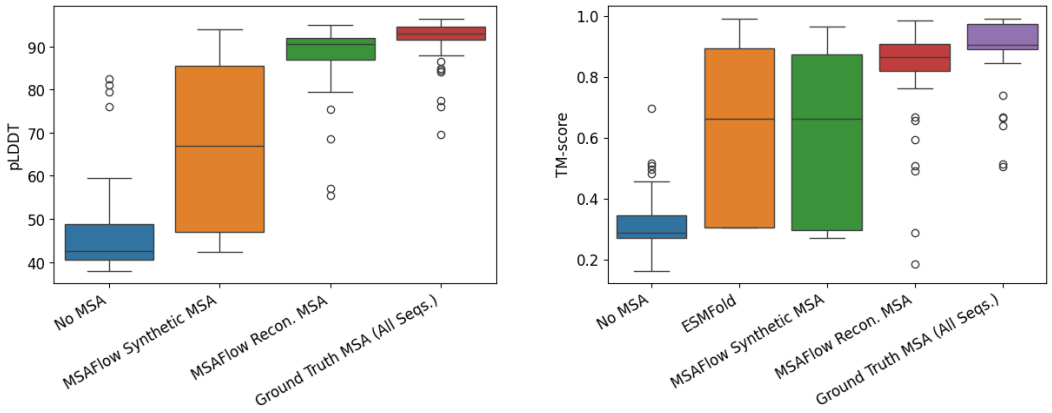

Figure 3: pLDDT and TM-scores for AF3 predictions of proteins from CAMEO with no MSA, the MSAFlow-based reconstructed MSA (32 sequences), a randomly subsampled 32-sequence MSA, and the ground truth deep MSA (approximately 7k sequences).

Furthermore, when attempting to build synthetic MSA embedding (i.e. MSAs generated via our latent diffusion model), we find that our decoder is able to reconstruct some signal from the generated MSA latents, achieving much higher quality than wihout using an MSA altogether, although the structure prediction accuracy remains worse than using the ground truth embedding itself. Another noteworthy point is that our model effectively compresses the heavy signal of full-depth evolutionary information encoded in thousands of aligned sequences into a single, fixed-size latent tensor that can be dynamically decoded into a range of sequences that remain evolutionarily related to the query. As a result, we keep almost all of the functional signal that matters for folding accuracy.

4.2 Augmenting shallow and single-sequence MSAs

We further evaluate our model on a dataset of sequences with limited evolutionary information derived from MSAGPT [6], which includes 200 proteins from CAMEO [14], CASP14, CASP15, and PDB [15] with either few or no sequences in their MSA (few-shot and zero-shot cases, respectively). For the zero-shot case, we embed the query sequence with ESM and use it as conditioning for our latent diffusion model, which generates a synthetic MSA embedding for the reference sequence. We generate embeddings over 10 different seeds and use low temperature sampling during the SDE forward pass for higher fidelity reconstructions, as detailed in [16]. We then decode 32 sequences from each of the 10 synthetic MSA embeddings and report the best pLDDT and TM-scores. We find

that our model significantly improves upon prior state-of-the-art MSA augmentation tools, which also seemed to yield poorer results when evaluated with AF3.

Table 1: The accuracy of MSAFlow-generated multiple sequence alignments compared to other state-of-the-art methods, as evaluated by AlphaFold3 protein structure prediction performance on a naturally scarce MSA dataset curated from CAMEO, PDB, and CASP.

	pLDDT		**TM-score**	
	Zero-shot	*Few-shot*	*Zero-shot*	*Few-shot*
No/Shallow MSA	73.1	**70.8**	0.55	0.58
EvoDiff (650M)	67.7	67.5	0.49	0.55
MSAGPT (3B)	71.6	70.3	0.53	0.58
MSAFlow (Ours,130M)	**75.2**	70.4	**0.62**	**0.60**

For the few-shot augmentation case, we use our latent flow matching model to generate synthetic embeddings for each sequence over 5 different seeds, and decode 32 sequences from each MSA embedding. We then decode 64 sequences from the ground-truth shallow MSA embedding and extract the 16 most diverse sequences across all generations, following [6]. We concatenate our generated sequences with the original shallow MSA and find that our model improves upon structure prediction accuracy for such cases.

4.3 Case Studies on *de novo* and intrinsically disordered proteins

We show that **MSAFlow** markedly improves structure prediction for notoriously difficult proteins by generating high-quality synthetic MSAs. We focus on three challenging cases from a sparse MSA dataset:

- **8B4K**: the N-terminal domain of Rfa1 complexed with a phosphorylated Ddc2 peptide—only 133 residues, with scarce evolutionary relatives.
- **8GI8**: a Rosetta-designed four-helix bundle with rigid backbone constraints, extraordinary thermal stability ($T_m > 90^\circ\mathrm{C}$), and NMR-validated topology (backbone RMSD = 1.11 Å).
- **8OKH**: the crystal structure of *Bdellovibrio bacteriovorus* Bd1399.

MSAFlow's synthetic MSAs significantly outperform both MSA-free predictions and those using MSAGPT, which lacks sufficiently precise coevolutionary signals. This highlights MSAFlow's strengths in addressing two key failure modes: (i) limited sequence homology and (ii) intrinsically flexible or disordered regions—by synthesizing information-rich, high-fidelity MSAs in latent space that modern folding models require.

4.4 Family-based Protein Design

To better demonstrate the strength of MSAFlow on few-shot generation and generalization to other downstream applications than AF3 prediction, we now provide new results on family-based enzyme design. **Our experiments demonstrate clear and significant advantages of MSAFlow, particularly for EC classes with limited sequences.** Following ProfileBFN [17], we generate sequences in a single shot using our model, for enzymes with less than 20 sequences in their corresponding EC class, using the sequences from the EC class as an MSA. We then use CLEAN [18] to determine their EC number, and compute the accuracy (i.e. how many generated designs match the ground truth EC number) and the uniqueness across all generated designs. We report the accuracy $\times$ uniqueness score as done by ProfileBFN, the current SOTA for this task. **MSAFlow exhibits SOTA performance on family-based enzyme design in both fixed and variable length settings.** Notably, ProfileBFN is confined to fixed-length generation, whereas MSAFlow learns a meaningful homology distribution that guides the placement of gaps, which effectively enables variable-length design with unprecedented success rate.

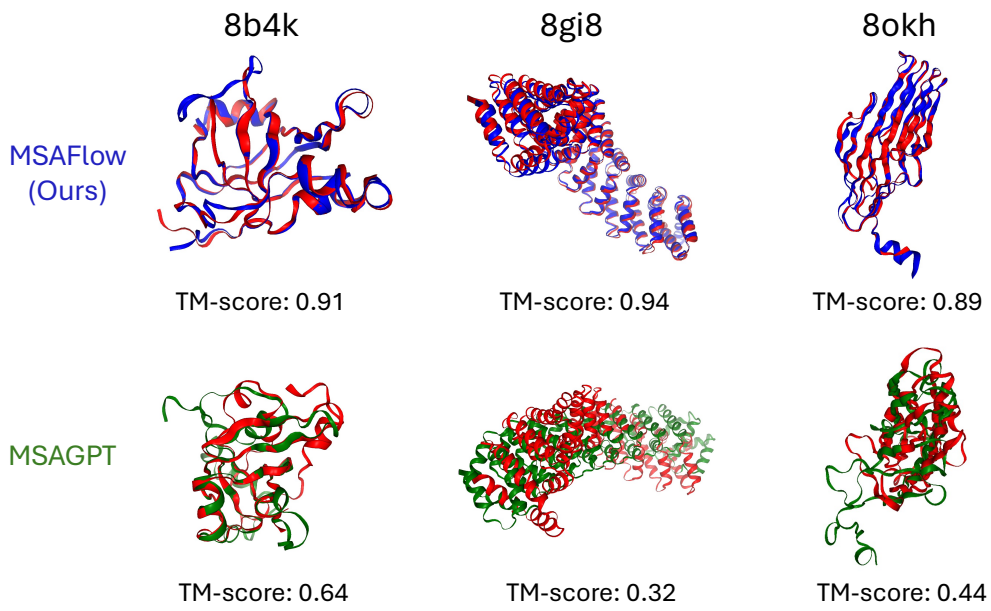

Figure 4: Visualization of improved structure prediction for zero-shot augmentation on de novo and disordered proteins with MSAFlow decoded synthetic MSAs, as compared to MSAs generated with MSAGPT. Blue represents predictions with an MSAFlow-generated MSA and green represents predictions with an MSAGPT-generated MSA. Red indicates the ground truth structure.

		Q15I65	**Q15BH7**	**P13280**	**P57298**
MSA Depth		15	12	13	15
# of Generated Sequences		1000	100	100	100
Accuracy × Uniqueness (Fixed)	EvoDiff	1.39%	0%	80%	5%
	ProfileBFN	42.67%	89%	**100%**	82%
	MSAFlow	**83.10%**	**84%**	**100%**	**95%**
Accuracy × Uniqueness (Variable)	EvoDiff	-	0%	0%	0%
	MSAGPT	-	35.59%	37.5%	24.98%
	MSAFlow	-	**92%**	**92%**	**84%**

Table 2: Performance comparison of MSAFlow with baseline methods on family-based enzyme design task across different EC classes.

5 Conclusion

MSAFlow integrates statistical flow matching with latent space optimization to enable bidirectional manipulation of multiple sequence alignments. By combining AlphaFold3-inspired permutation-equivariant embeddings with diffusion-based generation, it uniquely achieves both evolutionary signal compression and biologically plausible augmentation of sparse alignments. Comprehensive benchmarking across three critical applications—latent space reconstruction fidelity, shallow MSA augmentation for protein structure prediction, and synthetic alignment generation for underrepresented proteins—demonstrates MSAFlow's superiority, achieving state-of-the-art performance with only 130M parameters. MSAFlow's ability to generate evolutionarily coherent sequence ensembles creates new opportunities for designing orphan proteins and tackling de novo structure prediction challenges. Importantly, our framework also enables family-based design, where latent representations distilled from enzyme or protein families can guide the generation of sequences that remain faithful to family-level constraints while still exploring novel sequence diversity. Overall, MSAFlow advances both computational efficiency and conceptual modeling of protein sequence spaces through flow-based generation, paving the way for conditional protein engineering, resource-efficient applications, and family-level design of functional proteins.

References

[1] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175, February 2012.

[2] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017.

[3] L. Steven Johnson, Sean R. Eddy, and Elon Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1):431, August 2010.

[4] Liang Hong, Zhihang Hu, Siqi Sun, Xiangru Tang, Jiuming Wang, Qingxiong Tan, Liangzhen Zheng, Sheng Wang, Sheng Xu, Irwin King, Mark Gerstein, and Yu Li. Fast, sensitive detection of protein homologs using deep dense retrieval. *Nat. Biotechnol.*, pages 1–13, August 2024.

[5] Le Zhang, Jiayang Chen, Tao Shen, Yu Li, and Siqi Sun. MSA generation with seqs2seqs pretraining: Advancing protein structure predictions, 2024.

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

[7] Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex X. Lu, Nicolo Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need. Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2023/09/12/2023.09.11.556673.full.pdf.

[8] Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds, 2025.

[9] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, Sebastian W Bodenstein, David A Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B Fuchs, Hannah Gladman, Rishub Jain, Yousuf A Khan, Caroline M R Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024.

[10] ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix - advancing structure prediction through a comprehensive alphafold3 reproduction. *bioRxiv*, 2025.

[11] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J. O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M. Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M. Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Shiyang Chen, Minjia Zhang, Conglong Li, Shuaiwen Leon Song, Yuxiong He, Peter K. Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. Openfold: retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21(8):1514–1524, May 2024.

[12] Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. *arXiv preprint arXiv:2405.16441*, 2024.

[13] Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models, 2025.

[14] 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*, 86:387–398, March 2018.

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

[16] Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models, 2025.

[17] Jingjing Gong, Yu Pei, Siyu Long, Yuxuan Song, Zhe Zhang, Wenhao Huang, Ziyao Cao, Shuyi Zhang, Hao Zhou, and Wei-Ying Ma. Steering protein family design through profile bayesian flow. February 2025.

[18] Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, March 2023.

A Additional results

A.1 Zero-shot Prediction Comparison with ESMFold

Table 3: The accuracy of MSAFlow-generated multiple sequence alignments compared to ESMFold, evaluated on zero-shot protein structure prediction performance on a naturally scarce MSA dataset curated from CAMEO, PDB, and CASP.

	TM-score	
	Zero-shot	*Few-shot*
No/Shallow MSA	0.55	0.58
EvoDiff (650M)	0.49	0.55
MSAGPT (3B)	0.53	0.58
MSAFlow (Ours,130M)	**0.62**	**0.60**
ESMFold	0.58	NA

We further compared MSAFlow with ESMFold and other MSA generation models according to the TM-score after folding. The protein structure prediction research based on MSAFlow has demonstrated substantial results. Through evaluation on a naturally scarce MSA dataset, the results show that MSAFlow (applying only 130M parameters) achieved the optimal TM-scores in both zero-shot and few-shot scenarios, with scores of 0.62 and 0.60 respectively. In comparison, ESMFold scored 0.58 in zero-shot testing, while competing models such as EvoDiff (650M parameters) and MSAGPT (3B parameters) performed less effectively than MSAFlow. These results indicate MSAFlow's precise modeling in MSA generation and its computational efficiency.

A.2 Additional Case Studies

To further validate the robustness of MSAFlow's zero-shot predictions, we provide more cases for comparison. From the table 4, we can observe that MSAFlow achieves improvement on cases with different structural patterns as well as different families.

PDB ID	**Length**	**Description**	**GT**	**MSAGPT**	**MSAFlow**
6NW8_A	27	Scorpion venom toxin	0.39	0.40	**0.53**
6WKK_X	280	Phage capsid	0.28	0.27	**0.55**
7EQB_B	80	Central spindle assembly	0.65	0.58	**0.71**
7QRR_L	153	Noumeavirus	0.31	0.61	**0.83**
7ZOL_A	151	Cas 7-11 regulator	0.33	0.34	**0.67**

Table 4: Performance comparison of MSAFlow with baseline methods on clinically relevant proteins showing TM-Score improvements across different structural patterns and protein families.

A.3 Inference Speed and Memory Cost

In order to demonstrate that MSAFlow exhibits notable improvements in sampling efficiency compared to other MSA-based generative models, We benchmark MSAFlow against existing tools, attempting to generate 100 sequences conditioned on an existing MSA with 6 sequences on an NVIDIA A40 GPU, and observe the following:

	Latency Per Sequence	**Memory Consumption**
MSAFlow	**1.02s**	**5.8 GiB**
ProfileBFN	8.49s	7.7 GiB
MSAGPT	62.46s	41.6 GiB
EvoDiff	478.24s	4.0 GiB

Table 5: Sampling efficiency comparison of MSAFlow with baseline methods showing latency per sequence and memory consumption on NVIDIA A40 GPU for generating 100 sequences conditioned on an MSA with 6 sequences.

We find that MSAFlow has better sampling efficiency, both in terms of speed and memory. We can attribute this to the fact that our model only has to deal with L×H embedding of the MSA, rather than carry the quadratic cost of representing an MSA in the ambient space. The result shows that MSAFlow has the potential to be a highly light-weight and accurate MSA designer.

Moreover, our pipeline utilizes outputs from tools like MMseqs and HMMER for Multiple Sequence Alignment (MSA) reconstruction. A key advantage of this approach is its ability to generate high-quality MSAs even when these standard homology search methods fail to find sufficient homologous information. To provide a quantitative comparison of computational cost, we evaluated our MSAFlow model against HMMER and MMseqs2 for generating an MSA from a single query sequence (PDB 9BCZ_A from CAMEO, 644 amino acids). The empirical results are detailed below.

Method	**Wall Clock Time (s)**
MSAFlow (100 seqs)	**153.93**
HMMER	310.92
MMseqs2	497.73

Table 6: Computational cost comparison for generating MSA from query sequence alone (PDB 9BCZ_A from CAMEO, 644 AA) showing wall clock time in seconds.

These results show that MSAFlow achieves over 2× speedups compared to HMMER and MMseqs2, while still providing the ability to operate in settings where homology search fails. This confirms that MSAFlow not only addresses the coverage gap but also offers computational efficiency advantages over traditional methods.

A.4 Ablation Study of Reconstruction Sequences

We address using the additional ablation study on the reconstuction task with 2, 4, 8, 16, and 32 decoded MSA sequences, as well as the comparison with natural-MSA depth on 3 samples from the CAMEO reconstruction test set.

When we keep 2-4 sequences, the MSAFlow reconstructions beat the random ground-truth subsample. As we generate more sequences, the designed MSAs generally match that of the ground-truth samples (AlphaFold3 searched MSA), indicating that MSAFlow accurately captures structure patterns of protein families.

	PDB ID	**2**	**4**	**8**	**16**	**32**
Ground Truth Random Sample	9EJY	0.59	0.55	0.85	0.80	0.86
	9BIX	0.19	0.32	0.35	0.32	0.49
	9CVV	0.35	0.31	0.93	0.97	0.98
MSAFlow Reconstruction	9EJY	**0.61**	**0.61**	0.84	**0.83**	0.84
	9BIX	**0.28**	0.22	0.20	0.30	0.26
	9CVV	**0.43**	**0.62**	0.87	0.87	0.97

Table 7: Ablation study comparing MSAFlow reconstruction performance against ground truth random samples across different sequence counts on CAMEO reconstruction test set. Values represent performance metrics for MSA reconstruction quality. Numbers in the first row denotes the amounts of decoding MSA sequences.

A.5 Ablation Study on Synthetic and Reconstructed MSAs

The reconstruction pathway preserves the authentic signal from a limited, shallow MSA, while the latentflow pathway generates evolutionary diversity generalized from other MSA-rich proteins. These two tracks provide complementary signals that make the few-shot augmentation stronger. To provide evidence for this, we detail the separate contributions of each track below:

As shown in the table, the reconstruction path focuses on preserving crucial motif information within the limited observed sequences, which is reflected in the lower entropy signals in the shallow MSA. In contrast, the latentflow path generates synthetic MSAs that provide evolution-consistent diversity, resulting in higher entropy.

Few-shot task	TM Score	Avg Per-position Entropy
Syn-16	0.54	2.23
Rec-16	0.52	1.33
Syn+Rec-32	**0.57**	2.69
Syn+Rec+GT	**0.60**	2.58
MSAGPT+GT	0.58	1.33
GT	0.58	2.16

Table 8: Ablation study showing the complementary contributions of synthetic and reconstructed MSA pathways in few-shot tasks, demonstrating improved TM scores and entropy characteristics. **Syn** represents Synthetic MSAs; **Rec** represents Reconstructed MSAs. The number denotes amount of MSA sequences.

The combination of both tracks leads to an improvement in TM score and an increase in entropy. This observation confirms that the two tracks offer complementary signals, which synergistically improve quality. Finally, by augmenting the shallow ground truth MSA with the combined generation output, we improve prediction accuracy and achieve a better TM score than the MSAGPT baseline, which is what we report in Table 1. As can be seen, MSAFlow is the only method to achieve a better TM score than the ground truth, with an entropy value closest to it.

A.6 Ablation Study on ESM Embeddings

To clarify the individual contributions of the ESM embeddings and our proposed Statistical Flow-matching decoding mechanism, we provide ablation results for the MSAFlow zero-shot track trained to condition on the one-hot query sequence instead of the ESM embedding:

Method	TM Score
MSAGPT (3B)	0.53
MSAFlow Latent w/ one-hot (130M)	0.55
MSAFlow Latent w/ ESM2 (130M)	**0.62**

Table 9: Ablation study comparing the contribution of ESM embeddings versus one-hot sequence encoding in MSAFlow's zero-shot MSA augmentation performance.

The results demonstrate that the efficiency of our method. Moreover, ESM2 encoding provides more useful signals to address the evolutionary information.