DIFFUSION MSA PAIRFORMER: STRUCTURED MASKED DIFFUSION FOR PROTEIN REPRESENTATION AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple sequence alignments (MSAs) capture evolutionary constraints on protein structure and function through patterns of conservation and co-evolution. However, explicit modeling of MSA compositional structure—namely queries, homologous rows, and aligned columns—is notably lacking in current MSA language models. Relying on uniform token masking, existing approaches fail to fully exploit evolutionary information encoded within MSAs. We introduce Diffusion MSA Pairformer, a unified generative model that decomposes MSA diffusion into biologically motivated masking tasks and captures evolutionary dependencies through coupled MSA-pair representations. Empirically, this structured multi-task approach significantly enhances representation learning, with ablation studies confirming its crucial advantage over uniform masking. Our robust approach also enhances generation capabilities: in MSA augmentation and family-conditioned enzyme design, the model achieves superior performance by precisely capturing intra- and interchain structural constraints while balancing sequence diversity, structural consistency, and functional accuracy. These results establish structured MSA diffusion as a principled paradigm unifying evolutionary representation and protein design.

1 INTRODUCTION

Protein language models have achieved strong performance across protein understanding and design tasks (Alley et al., 2019; Rao et al., 2019; Rives et al., 2021; Elnaggar et al., 2022; Lin et al., 2023; Madani et al., 2023; Ferruz et al., 2022; Alamdari et al., 2026; Wang et al., 2024; Hayes et al., 2025). Through pretraining on billions of protein sequences, these models capture statistical patterns and evolutionary constraints that govern sequence variation (Rives et al., 2021; Lin et al., 2023), supporting applications in structure prediction, stability estimation, and functional annotation (Rao et al., 2019; Rives et al., 2021). Extending representation learning to generative modeling has further facilitated the design of novel proteins with desired properties, highlighting the utility of learned sequence distributions for protein engineering (Madani et al., 2023; Ferruz et al., 2022; Alamdari et al., 2026; Hayes et al., 2025).

Proteins evolve within homologous families under constraints imposed by structure and function (Sander & Schneider, 1991; Morcos et al., 2011). Multiple sequence alignments (MSAs) provide explicit evolutionary context through patterns of conservation and co-evolution that encode family-specific constraints on fold stability and functional specificity (Morcos et al., 2011; Rao et al., 2021). While sequence-based models learn evolutionary patterns implicitly through pretraining, MSA-based models leverage aligned homologs directly at inference time and model family-level evolutionary dependencies (Rao et al., 2021; Akiyama et al., 2026). This explicit evolutionary signal has driven advances in structure prediction (Jumper et al., 2021; Baek et al., 2021; Abramson et al., 2024; Lin et al., 2023) and mutation-effect prediction (Truong & Bepler, 2023; Notin et al., 2023).

Current MSA language models rarely excel in both strong representation and generation. MSA Transformer and MSA Pairformer achieve strong performance on contact prediction through masked language modeling (Rao et al., 2021; Akiyama et al., 2026), yet they are unable to generate sequences with MSA context. EvoDiff-MSA addresses this challenge through masked discrete diffusion, enabling MSA-conditioned sequence generation within a diffusion framework (Alamdari et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2026). However, EvoDiff-MSA employs uniform masking, similar to single-sequence diffusion models (Wang et al., 2024), without explicitly modeling the biological information in MSA or evaluating representation learning performance. MSAs contain distinct elements—query sequences, homologous contexts, and aligned columns—that may require differentiated design during training.

We present **Diffusion MSA Pairformer**, a generative MSA language model trained with multi-task masked diffusion. The approach decomposes MSA masking into biologically motivated subtasks that distinguish token-level, row-level, column-level, and query-level masking, and jointly supports representation learning, MSA augmentation, and family-conditioned enzyme generation.

Our contributions are:

- We introduce a structured multi-task masked-diffusion framework that unifies evolutionary representation learning and generation by explicitly modeling MSA structure.
- We show that Diffusion MSA Pairformer outperforms MSA Pairformer on mutation-effect, intrachain-contact, and protein–protein interface prediction. It also achieves competitive performance in MSA augmentation and family-conditioned enzyme generation.
- We reveal that pLDDT can mis-rank structures conditioned on generated MSAs, and develop a DMP-specific negative-log-likelihood(NLL) selector to improve candidate ranking.
- We construct a multi-Enzyme-Commission (EC) enzyme-design benchmark that extends conventional single-function evaluation by jointly measuring sequence diversity, EC-set recovery, global fold consistency, and ligand-aware active-site geometry.

2 Related Work

Homolog-conditioned protein models. Homologous-family context can be represented in several forms. Profile-based methods summarize aligned homologs as position-wise residue distributions, as in ProfileBFN, providing a compact representation of family-level conservation (Gong et al., 2025). Retrieval-augmented models, including Tranception, PoET, ProtMamba, and E1, incorporate homologous sequences without explicitly modeling residue-level alignment, using retrieval-based scoring or sequence-context architectures (Notin et al., 2022; Truong & Bepler, 2023; Sgarbossa et al., 2025; Jain et al., 2025). In contrast, MSA models preserve aligned residues and the associated row–column structure. MSA Transformer and EvoDiff-MSA apply axial attention to MSA tokens, whereas MSA Pairformer couples MSA and residue-pair representations to model evolutionary and structural dependencies (Rao et al., 2021; Alamdari et al., 2026; Akiyama et al., 2026).

MSA augmentation in protein structure prediction. MSA-dependent structure predictors require deep homologs, yet orphan proteins or de novo designed proteins often lack homology. Generative augmentation synthesizes additional MSA rows when retrieval yields shallow alignments. Approaches include sequence-to-sequence extension (Zhang et al., 2024), discrete diffusion (Alamdari et al., 2026), autoregressive decoding with AlphaFold2 guidance (Chen et al., 2024), profile-conditioned sampling (Gong et al., 2025), continuous embedding generation (Cao et al., 2025), and flow matching (Anonymous, 2026).

Family-conditioned protein generation. Family-conditioned generation seeks to sample novel protein sequences that remain consistent with the evolutionary constraints of a target family. Existing approaches condition autoregressive or diffusion-based generation on unaligned homolog sets, aligned MSAs, position-wise profiles, or learned family representations (Truong & Bepler, 2023; Alamdari et al., 2026; Sgarbossa et al., 2025; Gong et al., 2025; Truong & Bepler, 2025; Anonymous, 2026). Across these formulations, the central challenge is to explore diverse sequence space while preserving family-specific functional and structural constraints.

3 Preliminaries

3.1 MSA representation and conditional generation

Let $\mathcal{A}_{20}$ denote the canonical amino-acid alphabet and define the clean MSA vocabulary $\mathcal{V}_{\mathrm{MSA}} = \mathcal{A}_{20} \cup \{\mathtt{X}, \mathtt{B}, \mathtt{Z}, \mathtt{U}, \mathtt{O}, -\}$. The implementation additionally uses <pad> and <mask> as computational

special tokens. An MSA with depth N and aligned length L is represented as $\mathbf{X} = [x_{ij}] \in \mathcal{V}_{\text{MSA}}^{N \times L} = [\mathbf{q}; \mathbf{X}_{\text{ctx}}]$, where N is the number of sequences, L is the aligned length, $\mathbf{q}$ is the query sequence, and $\mathbf{X}_{\text{ctx}}$ contains the aligned homologs. Because the homolog rows have no intrinsic ordering, $[\mathbf{q}; \mathbf{X}_{\text{ctx}}]$ and $[\mathbf{q}; \mathbf{P}\mathbf{X}_{\text{ctx}}]$ represent the same MSA context for any permutation matrix $\mathbf{P}$ acting on the homolog rows.

For a generation index set $\mathcal{G} \subseteq [N] \times [L]$, let $\mathbf{X}_{\mathcal{G}}$ denote the tokens to be generated and $\mathbf{X}_{\overline{\mathcal{G}}}$ the observed complement. Diffusion MSA Pairformer, parameterized by θ, models

$$p_\theta\left(\mathbf{X}_{\mathcal{G}} \mid \mathbf{X}_{\overline{\mathcal{G}}}, \mathcal{G}\right). \tag{1}$$

Query generation corresponds to $\mathcal{G} = \{1\} \times [L]$. Homolog-row generation uses $\mathcal{G} = \mathcal{R} \times [L]$ for a selected row set $\mathcal{R} \subseteq \{2, \ldots, N\}$, including the multi-row setting used for MSA augmentation. The corresponding masking objectives and decoding procedure are specified in Section 4.2.

4 METHODS

4.1 STRUCTURED MASKED DIFFUSION OVER MSAS

Our model extends the MSA Pairformer architecture (Akiyama et al., 2026) with a structured mask diffusion framework. The architecture maintains an MSA representation $\mathbf{H} \in \mathbb{R}^{N \times L \times d_{\text{msa}}}$ and pair representation $\mathbf{Z} \in \mathbb{R}^{L \times L \times d_{\text{pair}}}$, coupled through query-based outer product and triangle multiplication update operations. We extend the model with masked discrete diffusion (Austin et al., 2021; Sahoo et al., 2024), which models a continuous family of noise levels rather than a fixed masking distribution. Crucially, we exploit the compositional structure of MSAs by defining task-specific masking schemes that mask tokens, rows, columns, and query-conditioned contexts at varying rates, enabling the model to learn both local token dependencies and global alignment covariation.

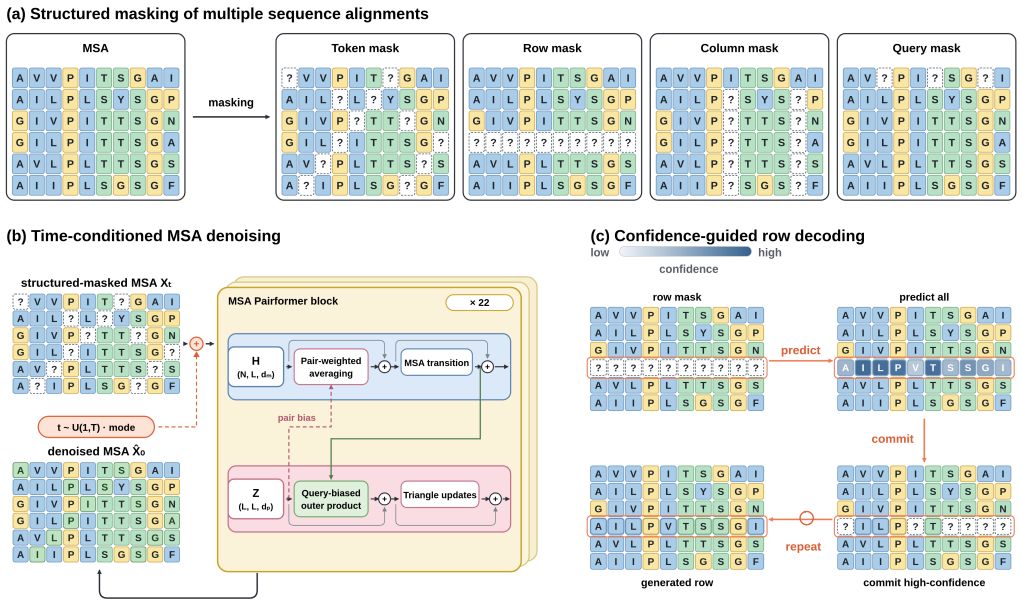

Figure 1: **Diffusion MSA Pairformer (DMP). (a)** Structured multi-task masking applies token-level, row-level, column-level, and query-level corruption to an MSA. **(b)** The masked MSA is processed by the MSA Pairformer backbone during training. **(c)** Conditional generation uses iterative confidence-ordered decoding: high-confidence tokens are committed, while unresolved low-confidence positions remain masked for later steps.

The masking process replaces selected valid MSA tokens with <mask>. Let T denote the number of diffusion timesteps. For $t \in \{1, \dots, T\}$, we set $\rho_t = t/T$ and apply a task-specific binary mask $\mathbf{M}_t$:

$$[\mathbf{X}_t]_{ij} = \begin{cases} \texttt{<mask>}, & M_{t,ij} = 1, \\ x_{ij}, & M_{t,ij} = 0. \end{cases} \tag{2}$$

The denoiser predicts the clean token x_{ij} from $(\mathbf{X}_t, t)$.

4.2 Multi-task masking

Let $\mathcal{I}_q = \{1\}, \mathcal{I}_h = \{2, \dots, N\}$, and $\mathcal{J} = \{1, \dots, L\}$ denote query row, homolog rows, and column indices. Diffusion time t is sampled uniformly from $\{1, \dots, T\}$, and the mask ratio $\rho_t = t/T$ controls masking intensity, with $T = 500$. The four tasks are:

Token diffusion. Valid positions are masked independently, $M^{\text{tok}}_{t,ij} \sim \text{Bernoulli}(\rho_t)$, providing token-level denoising supervision across training tasks.

Row diffusion. We sample $\mathcal{S}^{\text{row}}_t \subseteq \mathcal{I}_h$ with $|\mathcal{S}^{\text{row}}_t| = \text{round}(\rho_t |\mathcal{I}_h|)$ and set $M^{\text{row}}_{t,ij} = \mathbb{1}\{i \in \mathcal{S}^{\text{row}}_t\}$. This trains the model to generate new homolog sequences conditioned on existing family members.

Column diffusion. We sample $\mathcal{S}^{\text{col}}_t \subseteq \mathcal{J}$ with $|\mathcal{S}^{\text{col}}_t| = \text{round}(\rho_t \min(L, C_{\max}))$ and set $M^{\text{col}}_{t,ij} = \mathbb{1}[i \in \mathcal{I}_h]\mathbb{1}[j \in \mathcal{S}^{\text{col}}_t]$ where $C_{\max}$ limits the maximum number of columns that can be masked. This trains reconstruction of masked alignment columns from the remaining MSA context, which contributes to co-evolutionary learning.

Query diffusion. We sample $\mathcal{S}^{q}_t \subseteq \mathcal{J}$ with $|\mathcal{S}^{q}_t| = \text{round}(\rho_t L)$ and set $M^{q}_{t,ij} = \mathbb{1}[i = 1]\mathbb{1}[j \in \mathcal{S}^{q}_t]$. This directly trains family-conditioned protein generation, enabling sampling and scoring of novel sequences consistent with a given homolog alignment.

4.3 Training pipeline

We use a maximum training shape of $N = 128, L = 512$ as our MSA shape in training. Training proceeds in two stages. The first epoch uses token-level diffusion as warm-up. Subsequently, one task is sampled per minibatch from a task mixture $\boldsymbol{\pi} = (\pi_{\text{token}}, \pi_{\text{row}}, \pi_{\text{col}}, \pi_{\text{query}})$. Let $\mathcal{M}_t = \{(i, j) : M_{t,ij} = 1\}$ denote masked positions. The training objective minimizes weighted cross-entropy over masked positions:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \mathbf{X}_0, \mathbf{M}_t} \left[\frac{\sum_{(i,j) \in \mathcal{M}_t} w_{ij} \cdot \ell_{\text{CE}}(f_\theta(\mathbf{X}_t, t)_{ij}, x_{ij})}{\sum_{(i,j) \in \mathcal{M}_t} w_{ij}} \right], \tag{3}$$

where $f_\theta(\mathbf{X}_t, t)_{ij}$ are predicted logits at position (i, j), and w_{ij} are position-dependent weights. Exact task probabilities, position weights, warm-up schedules, and optimizer settings are given in Appendix A.2.

4.4 Inference and decoding

At inference, positions in the generation set $\mathcal{G}$ (defined in Section 3.1) are initialized to <mask>, while observed positions $\overline{\mathcal{G}}$ remain fixed. We employ iterative confidence-ordered decoding (Chang et al., 2022) for S reverse steps. At reverse step $r \in \{S, \dots, 1\}$, the model predicts the unresolved positions at diffusion time t_r. Temperature-scaled softmax with temperature τ defines the categorical sampling distribution, and confidence at position (i, j) is $\gamma_{ij} = \max_v p_\theta(v \mid \mathbf{X}^{(r)}, t_r)_{ij}$. Let m_r be the number of unresolved positions at step r. The implementation commits the highest-confidence positions according to

$$t_r = \max\Big(1, \text{round}\Big(\frac{r}{S} T\Big)\Big), \qquad k_r = \min\Big\{m_r, \max\Big(1, m_r - \text{round}\Big(\frac{r-1}{S} m_r\Big)\Big)\Big\}, \tag{4}$$

where k_r positions with highest confidence are sampled and committed, while unresolved positions remain masked. At $r = 1$, all remaining positions are committed. Decoding hyperparameters, including τ and S, are specified in Appendix A.4.

5 RESULTS

5.1 STRUCTURED MULTI-TASK DIFFUSION IMPROVES EVOLUTIONARY REPRESENTATIONS

Structured multi-task masking is designed to model the compositional structure of MSAs, enabling richer evolutionary representations than uniform masking. To evaluate whether this paradigm improves protein understanding, we assess the learned representation on three complementary benchmarks: mutation-effect prediction (Notin et al., 2023), intrachain contact prediction (Rao et al., 2021; Akiyama et al., 2026), and protein–protein interaction (PPI) interface prediction (Akiyama et al., 2026). Mutation-effect scoring probes whether the model captures sequence-level functional constraints; contact prediction evaluates geometric modeling within single chains; interface prediction extends this to interchain structural dependencies. Task definitions and evaluation protocols are provided in Appendix B.1.

Diffusion MSA Pairformer consistently outperforms MSA Pairformer across all three benchmarks (Figure 2). On ProteinGym, Spearman ρ improves from 0.433 to 0.449, surpassing MSA Transformer and the evaluated ESM2 and ESM-C baselines. Our model also improves the mean long-range $P@L$ from 0.502 to 0.559 on the CASP15 intrachain benchmark, indicating stronger geometry modeling from MSAs. On the 27-complex PPI benchmark, median interface $P@K$ improves from 0.521 to 0.538, showing that the benefits of structured multi-task diffusion extend to interchain contact prediction. All three benchmarks indicate that multi-task mask diffusion improves protein representations from evolutionary context.

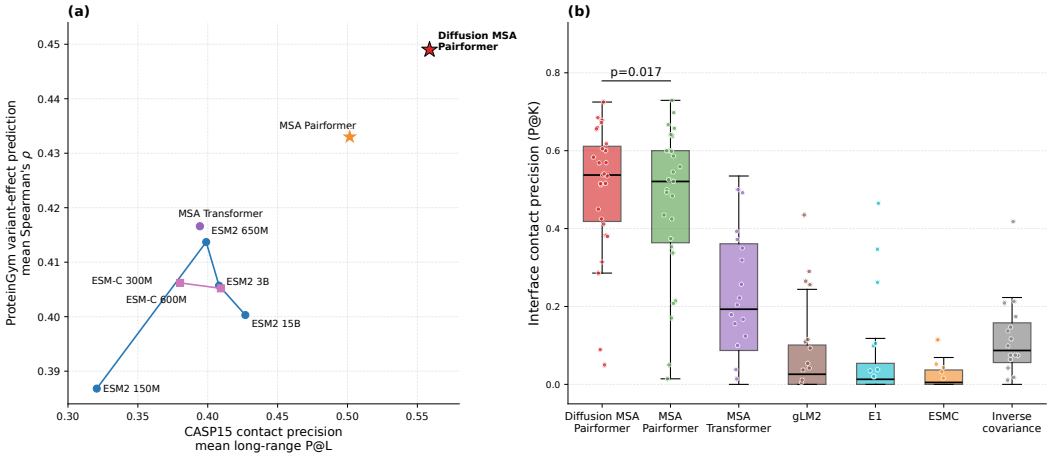

Figure 2: **Representation benchmarks. (a)** ProteinGym weighted-mean Spearman's ρ versus CASP15 mean long-range $P@L$ across 57 targets. All contact values on the horizontal axis are obtained from the same 57-target evaluation. **(b)** PPI interface $P@K$ across 27 complexes. Data provenance is provided in Appendix B.1.

5.2 ABLATION STUDY OF STRUCTURED MULTI-TASK MASKING

To identify the contribution of each structured masking task to representation learning, we conduct ablation studies with MSAs truncated to $N = 32$. We compare the full structured mixture with uniform token masking and three variants that remove query-, row-, or column-level masking, while reallocating the removed probability to uniform masking. Detailed experimental procedures and statistical methods are provided in Appendix B.3.

Prior protein diffusion models typically employ uniform token masking for training (Alamdari et al., 2026). Our uniform-mask ablation reveals that the full structured baseline significantly outperforms this conventional approach on all three benchmarks (Figure 3): ProteinGym ρ drops from 0.442 to 0.421, PPI median $P@K$ from 0.471 to 0.297, and CASP15 long-range $P@L$ from 0.531 to 0.485 (all $p < 0.0001$). These controlled comparisons show that the representation gains arise from the structured task mixture rather than from diffusion alone.

Figure 3 further resolves the contribution of each structured task to representation learning. Removing query, row, or column masking significantly reduces CASP15 contact prediction. Removing column masking also reduces ProteinGym and PPI performance, whereas query removal affects PPI prediction

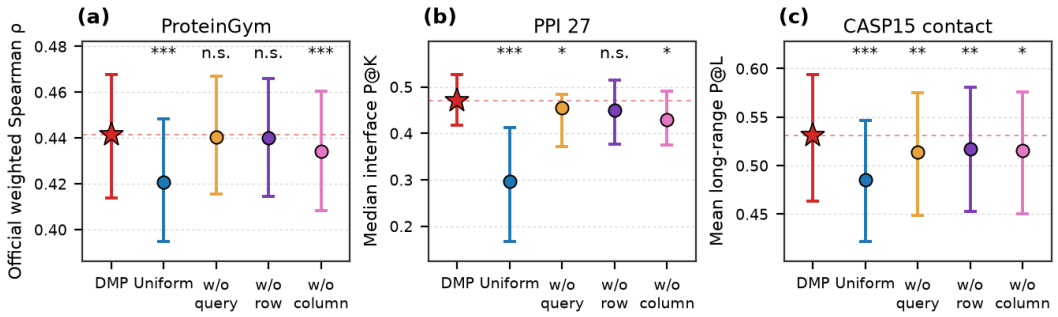

Figure 3: **Contributions of the three structured masking tasks.** $N32$ ablation setup. Points and error bars show point estimates with 95% bootstrap confidence intervals; the dashed line marks the full model. Stars mark statistically significant paired differences from the full model. Full test definitions and statistics are reported in Appendix B.3.

and row removal does not reach significance on PPI. These controlled removals demonstrate that each masking task makes a distinct contribution to the learned representation: no single task is redundant, as removing any one degrades performance on at least one benchmark. These results thus confirm that structured masking is not merely a reweighting of uniform masking but a necessary decomposition of MSA structure into complementary learning tasks. Exact test statistics are reported in Appendix B.3.

5.3 MSA AUGMENTATION ON ORPHAN PROTEIN STRUCTURE PREDICTION

Protein structure prediction from MSAs benefits from abundant evolutionary information (Jumper et al., 2021; Lin et al., 2023), but many proteins from viral, metagenomic, or poorly characterized families have few detectable homologs. MSA-generation methods synthesize homologs intended to augment this limited context (Chen et al., 2024; Gong et al., 2025).

Setup. We adopt the orphan-protein subset from Chen et al. (2024): 199 proteins whose native MSAs are less than 20 members. We further filter by retaining only cases where ColabFold (Mirdita et al., 2022) prediction still achieves TM-score below 0.5, producing a 52-case Hard52 subset. For each case, we generate 40 candidate MSAs and submit to ColabFold or ESMFold2. We report pLDDT top-1 TM, the DMP pLDDT–NLL-selected TM, and oracle TM (the retrospective maximum TM-score over each candidate pool). Complete protocols are in Appendix C.1.

Results on AlphaFold2. Table 1 shows that all evaluated MSA-generation methods achieve oracle TM-scores exceeding both AF2 baselines, with DMP reaching the highest oracle performance. For each generator, the oracle exceeds the pLDDT top-1 result, consistent with evidence that deeper MSAs can inflate pLDDT without corresponding structural improvements (Akdel et al., 2022; Terwilliger et al., 2024). MSAGPT escapes this failure by generating higher gap content or more conservative sequence and obtains the highest pLDDT top-1 TM-score in this comparison (Figure 6). To address DMP's ranking gap, we evaluate a DMP-specific selector that combines mean pLDDT with the conditional NLL assigned by DMP to the candidate's added MSA tokens (Appendix C.2). Lower conditional NLL indicates that the added rows are more compatible with the fixed raw MSA. It raises the selected TM-score from 0.3268 to 0.3581, the highest selected TM-score in the AF2 comparison. MSA composition analysis and structure prediction case studies are in Appendix C.4.

Generalization to ESMFold2. Single-sequence protein language models achieve strong structure prediction performance by distilling evolutionary constraints from large-scale pretraining (Candido et al., 2026). To assess whether MSA generation provides information beyond what sequence models capture, Table 2 evaluates whether generated MSAs improve ESMFold2 predictions on the same Hard52 cases. DMP yields the highest oracle TM-score, exceeding both single-sequence ESMFold2 and the shallow-MSA condition. It also surpasses MSAGPT and ProfileBFN on oracle TM-score and selected TM-score. These results demonstrate that DMP-generated MSAs enhance structural signals on either MSA-based AF2 or pLM-based ESMFold2, as reflected by the oracle results and the DMP-specific conditional-NLL selector. Representative AlphaFold2 and ESMFold2 predictions are shown in Figures 7 and 8, respectively.

Table 1: **MSA generation for orphan-protein structure prediction (Hard52 subset).** pLDDT top-1 TM uses mean pLDDT to rank all the samples. Oracle TM is the retrospective maximum across 40 samples. AF2 full: AlphaFold2 with shallow MSA. AF2 poor: AlphaFold2 with query-only input. AF2 full win rate is the fraction of cases where the selected prediction exceeds the AF2 full baseline. **Bold**: best; underline: second-best.

Method	pLDDT top-1 TM ↑	pLDDT ↑	Oracle TM ↑	AF2 full win rate ↑
AF2 full	0.3466	51.76	0.3780	0/52 (0.0%)
AF2 poor	0.3227	50.93	0.3633	17/52 (32.7%)
ProfileBFN	0.3023	57.11	0.3805	14/52 (26.9%)
MSAGPT	0.3490	57.98	0.4157	21/52 (40.4%)
DMP (pLDDT top-1)	0.3268	**61.40**	**0.4178**	21/52 (40.4%)
DMP (pLDDT–NLL selector)	**0.3581**	—	**0.4178**	**28/52(53.8%)**

Table 2: **ESMFold2 transfer evaluation (Hard52 subset).** Selected TM is based on mean-pLDDT ranking for all rows except the DMP pLDDT–NLL selector row, which uses the Appendix C.2 score. ESMFold2 win rate is the fraction of cases where the reported selected prediction has higher TM-score than the ESMFold2 single-sequence baseline. **Bold**: best; underline: second-best.

Method	pLDDT top-1 TM ↑	pLDDT ↑	Oracle TM ↑	ESMFold2 win rate ↑
ESMFold2	0.4571	60.47	0.5092	0/52 (0.0%)
ESMFold2 + shallow-MSA	0.4589	65.21	0.5083	**25/52(48.1%)**
ProfileBFN	0.3976	64.62	0.4826	11/52 (21.2%)
MSAGPT	0.4572	67.03	0.5207	24/52 (46.2%)
DMP (pLDDT top-1)	0.4440	**67.17**	**0.5312**	23/52 (44.2%)
DMP (pLDDT–NLL selector)	**0.4606**	—	**0.5312**	**25/52(48.1%)**

5.4 Family-Conditioned Enzyme Design and Evaluation

Family-conditioned generation aims to preserve evolutionary constraints associated with enzymatic function. Prior evaluations have focused primarily on mono-functional families (Gong et al., 2025; Truong & Bepler, 2023), whereas many enzymes carry multiple catalytic annotations or exhibit activity toward multiple reactions or substrates (Khersonsky & Tawfik, 2010; Tawfik, 2010; Copley, 2012). Such activity contributes to the evolution of new enzyme functions and is relevant to biocatalysis and pathway engineering (Copley, 2012; Humble & Berglund, 2011; Carbonell & Faulon, 2010). Multi-functional family design is therefore a stricter test of whether generated sequences retain several family-level functional constraints.

We evaluate family-conditioned generation under progressively stronger constraints. Single-enzyme design tests preservation of a family's annotated EC label, whereas multi-EC design requires simultaneous recovery of multiple enzymatic annotations. Full enzyme design further restricts the multi-EC benchmark to families with experimental ligand-bound structures, allowing us to assess EC-set recovery together with global fold consistency and ligand-aware active-site geometry.

5.4.1 Single-enzyme design: mono-EC preservation

We adopt the six-family benchmark reported with ProfileBFN and MSAFlow (Gong et al., 2025; Anonymous, 2026), spanning deep and shallow MSAs, and compare Diffusion MSA Pairformer against EvoDiff-MSA, ProfileBFN, and MSAFlow, which are advanced homolog-based generators. Functional accuracy is scored with CLEAN EC-set prediction (Yu et al., 2023), following Appendix D.1.

Table 3 shows that Diffusion MSA Pairformer attains near-ceiling accuracy on the two deep-MSA families and is strongest or a close second on the shallow-MSA cases. EvoDiff-MSA performs less favorably on these shallow families, while ProfileBFN and MSAFlow remain competitive on subsets of cases. Overall, DMP remains competitive with prior family-conditioned generators in the mono-EC setting, providing a strong baseline for the stricter multi-EC and structure-ligand evaluations below.

Table 3: **Single-enzyme design benchmark.** We report Acc. × Uniqueness (↑), measuring the fraction of generated sequences that are both functionally accurate and unique. **Bold**: best; underline: second-best.

	P40925	Q7X7H9	Q15165	Q15BH7	P13280	P57298
MSA Depth	572	443	15	12	13	15
# Generated Seqs	1000	1000	1000	100	100	100
Acc. × Uniqueness						
EvoDiff	27.9%	88.7%	1.4%	0%	80%	5%
ProfileBFN	95.2%	99.0%	42.7%	**89%**	**100%**	82%
MSAFlow	—	—	83.1%	84%	**100%**	**95%**
Diffusion MSA Pairformer	**100%**	**99.9%**	**91.3%**	77.1%	96.4%	91.3%

5.4.2 Multi-EC enzyme design: functional accuracy and diversity

We additionally cluster 15 families from the whole CLEAN dataset in which every member carries at least two EC labels (Appendix D.2). CLEAN predicts an EC set for each generated sequence. Any-EC accuracy evaluates overlap with the family EC set, whereas all-EC accuracy evaluates recovery of the complete annotated EC set.

Table 4 shows that Diffusion MSA Pairformer attains the highest all-EC accuracy with diverse sequences, demonstrating that the model effectively learns functional constraints within multi-EC families. ProfileBFN achieves comparable any-EC accuracy but lower all-EC accuracy, indicating that it captures individual enzymatic activities but struggles to satisfy all functional requirements simultaneously. PoET provides the closest competing balance between all-EC recovery and uniqueness. EvoDiff-MSA produces entirely unique samples in this evaluation but recovers the complete EC set less frequently than the other family-conditioned generators.

This result is consistent with a role for coupled residue dependencies in multi-functional enzyme families (Khersonsky & Tawfik, 2010; Jeon et al., 2011; Tawfik, 2010). The triangle multiplication in Pairformer and multi-task masking are consistent with such dependencies, whereas profile representations retain position-wise marginals.

Table 4: **Multi-EC enzyme design benchmark.** Results for 15 families with multiple EC annotations. Any-EC, all-EC, and Unique are means computed over all 1000 generated samples. Any×U and All×U are uniqueness-adjusted metrics computed per sample and then averaged. Pairwise identity: mean identity among generated sequence pairs (↓). **Bold**: best; underline: second-best.

Method	Unique ↑	Any-EC ↑	All-EC ↑	Any×U ↑	All×U ↑	Pairwise identity ↓
EvoDiff-MSA	**1.000**	0.531	0.472	0.531	0.472	—
PoET	0.998	0.901	0.859	0.899	0.858	—
ProfileBFN	**1.000**	0.931	0.818	**0.931**	0.818	**0.546**
DMP (Ours)	0.955	**0.933**	**0.925**	0.889	**0.881**	0.567

5.4.3 Structure-ligand design: fold and active-site compatibility

CLEAN-based EC-set prediction is sequence-level and cannot establish fold or active-site structure. We therefore restrict the multi-EC datasets to six families with multiple ligand-bound crystal structures, spanning single-pocket and dual-domain architectures (Appendix D.4), and evaluate global fold and ligand-aware active-site geometry.

Table 5 reports protein-only evaluation, aligning ESMFold2-predicted structures (Candido et al., 2026) to the reference crystals. With the exception of EvoDiff-MSA, whose EC accuracy remains poor on this subset as in Table 4, all methods produce reasonably well-folded, functionally consistent protein structures, and Diffusion MSA Pairformer leads on both fold accuracy and all-EC accuracy simultaneously. ProfileBFN's structures are comparatively less accurate, and PoET's all-EC accuracy trails DMP's despite similar fold quality.

For Table 6, ESMFold2 co-folds each generated protein with the organic ligand(s) specified for the corresponding experimental state (Appendix D.3). We use the FoldBench ligand-success criterion: symmetry-corrected, binding-site-superposed LRMSD < 2 Å and lDDT-PLI > 0.8 (Xu et al., 2026). Table 6 reports family-macro success rates and continuous ligand-, pocket-, and fold-level

scores across the six families. DMP has the highest mean success rate and the best family-macro LRMSD, lDDT-PLI, and lDDT-LP. This convergent evidence across functional and structural criteria establishes that DMP-generated sequences satisfy the coupled evolutionary constraints inherent to multi-functional enzyme families, where a single scaffold must accommodate multiple catalytic activities without compromising structural integrity.

Table 5: **Full enzyme design: protein-only evaluation.** Results for six complex multi-EC families. TM: mean TM-score (↑). Any-EC/All-EC: functional accuracy (↑). Pairwise identity: mean identity among generated sequence pairs (↓). Unique: fraction distinct (↑). **Bold**: best; underline: second-best.

Method	pLDDT ↑	TM ↑	Any-EC ↑	All-EC ↑	Pairwise identity ↓	Unique ↑
EvoDiff-MSA	72.91	0.818	0.400	0.342	—	**1.000**
PoET	90.85	0.903	0.958	0.887	—	0.997
ProfileBFN	85.92	0.901	0.993	0.804	**0.55**	**1.000**
DMP	**92.35**	**0.946**	**1.000**	**0.929**	0.56	0.962

Table 6: **Full enzyme design: ligand-aware co-fold evaluation.** Family-macro results across six families and 16 ligand-bound states, with 1,000 candidates per state. LRMSD: ligand-pose deviation (↓); lDDT-PLI: protein–ligand distance recovery (↑); lDDT-LP: local pocket score (↑); TM: global protein fold similarity (↑). **Bold**: best; underline: second-best.

Method	Success (%) ↑	LRMSD (Å) ↓	lDDT-PLI ↑	lDDT-LP ↑	TM ↑
EvoDiff-MSA	0.5	10.22	0.173	0.215	0.818
PoET	14.5	7.78	0.266	0.307	0.903
ProfileBFN	31.1	6.51	0.542	0.602	0.901
DMP	**42.2**	**4.83**	**0.633**	**0.674**	**0.945**

6 Discussion

Our results support structured multi-task masking as a common training basis for evolutionary representation learning and family-conditioned generation. The ablations show task-dependent contributions from the four masking tasks, while the full mixture outperforms uniform masking across the evaluated representation benchmarks showing the significance of explicit modeling of MSA. In MSA augmentation and family-conditioned enzyme design tasks, DMP achieves the highest oracle TM-score performance and best function-structure consistency, demonstrating that structured multi-task masking transfers evolutionary dependencies from representation learning into generation. These results also suggest that modeling MSA can learn co-evolution and covariance constraints explicitly to improve protein understanding, which may be more effective than profile-based or retrieval-based methods.

7 Limitations

Although we achieve promising oracle TM-score improvement in MSA augmentation, pLDDT inflation and mis-ranking remain challenges for selecting high-quality candidates. Our selector shows potential of sampling-based adjustment on pLDDT, but better confidence scorer is needed in this field. The multi-EC benchmark contains 15 eligible families, with ligand-aware evaluation for six families. Expanding family coverage and the benchmark with additional curated multifunctional enzymes, structures, and experiments would evaluate generalization more broadly.

AI USE STATEMENT

In this work, we used generative AI tools to design and provide feedback on research methodology and experiments, implement methods, assist with translation, and clean and reformat datasets.

We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, support qualitative or thematic data analysis, or interpret results. Providing critical ingredients for proofs and assisting in the writing of proofs are not applicable to this work.

Additionally, we used generative AI tools to suggest experimental parameters, create and edit software code, draft parts of the manuscript, summarize existing literature, edit the text for readability, format references, and suggest the paper structure. We reviewed all AI-assisted method suggestions, code, data-processing steps, translations, and manuscript text; all experiments were run and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Lukas Jarosch, Dan Berenberg, Ian Fisk, Andrew M. Watkins, Stephen Ra, Richard Bonneau, and Mohammed AlQuraishi. OpenProteinSet: Training data for structural biology at scale. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0204. Datasets and Benchmarks Track.

Mehmet Akdel, Douglas E. V. Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O. Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L. Good, Roman A. Laskowski, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Petras Kundrotas, Victoria Ruiz Serra, Carlos H. M. Rodrigues, Alistair S. Dunham, David Burke, Neera Borkakoti, Sameer Velankar, Adam Frost, Jérôme Basquin, Kresten Lindorff-Larsen, Alex Bateman, Andrey V. Kajava, Alfonso Valencia, Sergey Ovchinnikov, Janani Durairaj, David B. Ascher, Janet M. Thornton, Norman E. Davey, Amelie Stein, Arne Elofsson, Tristan I. Croll, and Pedro Beltrao. A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, 2022. doi: 10.1038/s41594-022-00849-w.

Yo Akiyama, Zhidian Zhang, Olivia Tang, Rachel Seongeun Kim, Milot Mirdita, Martin Steinegger, and Sergey Ovchinnikov. Expanding the scope of protein language modeling to protein-protein interactions with MSA Pairformer. *Cell*, 189(16):4964–4979.e8, 2026. doi: 10.1016/j.cell.2026.06.029.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Robert Strome, Alan M. Moses, Alex X. Lu, Nicolò Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need. *eLife*, 2026. doi: 10.7554/eLife.112029. URL https://elifesciences.org/reviewed-preprints/112029. eLife reviewed preprint.

Ethan C. Alley, Grigory Khimulya, Siddhartha Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16:1315–1322, 2019. doi: 10.1038/s41592-019-0598-1.

Anonymous. MSAFlow: A unified approach for MSA representation, augmentation, and family-based protein design. Anonymous OpenReview preprint submitted to ICLR 2026, 2026. URL https://openreview.net/pdf?id=ZWhE0ObivK. The consulted PDF identifies itself as under double-blind review.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754.

M. Biasini, T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippsen, and T. Schwede. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69 (5):701–709, 2013. doi: 10.1107/S0907444913007051.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735. URL https://www.biorxiv.org/content/10.64898/2026.06.03.729735v1. Preprint.

Hanqun Cao, Xinyi Zhou, Zijun Gao, Chenyu Wang, Xin Gao, Zhi Zhang, Chunbin Gu, Ge Liu, and Pheng-Ann Heng. Lightweight MSA design advances protein folding from evolutionary embeddings. *arXiv preprint arXiv:2507.07032*, 2025. URL https://arxiv.org/abs/2507.07032. Preprint.

Pablo Carbonell and Jean-Loup Faulon. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics*, 26(16):2012–2019, 2010. doi: 10.1093/bioinformatics/btq317.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022. doi: 10.1109/CVPR52688.2022.01104.

Bo Chen, Zhilei Bei, Xingyi Cheng, Pan Li, Jie Tang, and Le Song. MSAGPT: Neural prompting protein structure prediction via MSA generative pre-training. In *Advances in Neural Information Processing Systems*, volume 37, pp. 37504–37534, 2024. doi: 10.52202/079017-1184.

Shelley D. Copley. Toward a systems biology perspective on enzyme evolution. *Journal of Biological Chemistry*, 287(1):3–10, 2012. doi: 10.1074/jbc.R111.254714.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Noelia Ferruz, Steffen Schmidt, and Birte Hocker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, 2022. doi: 10.1038/s41467-022-32007-7.

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. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/83c54476af83ec99266820a842e201c2-Abstract-Conference.html.

Thomas Hayes, Roshan Rao, Halil Akın, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

Maria Svedendahl Humble and Per Berglund. Biocatalytic promiscuity. *European Journal of Organic Chemistry*, (19):3391–3401, 2011. doi: 10.1002/ejoc.201001664.

Sarthak Jain, Joel Beazer, Jeffrey A. Ruffolo, Aadyot Bhatnagar, and Ali Madani. E1: Retrieval-augmented protein encoder models. *bioRxiv*, 2025. doi: 10.1101/2025.11.12.688125. Preprint.

Jouhyun Jeon, Hyun-Jun Nam, Yoon Sup Choi, Jae-Seong Yang, Jihye Hwang, and Sanguk Kim. Molecular evolution of protein conformational changes revealed by a network of evolutionarily coupled residues. *Molecular Biology and Evolution*, 28(9):2675–2685, 2011. doi: 10.1093/molbev/msr094.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, August Zidek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Olga Khersonsky and Dan S. Tawfik. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Review of Biochemistry*, 79:471–505, 2010. doi: 10.1146/annurev-biochem-030409-143718.

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. doi: 10.1126/science.ade2574.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html.

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. doi: 10.1093/nar/gkw1081.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: Making protein folding accessible to all. *Nature Methods*, 19:679–682, 2022. doi: 10.1038/s41592-022-01488-1.

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, Jose N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011. doi: 10.1073/pnas.1111471108.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N. Gomez, Debora Marks, and Yarin Gal. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16990–17017, 2022. URL https://proceedings.mlr.press/v162/notin22a.html.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, 2023. doi: 10.52202/075280-2810. Datasets and Benchmarks Track.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9689–9701, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8844–8856. PMLR, 2021. URL https://proceedings.mlr.press/v139/rao21a.html.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander M. Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184, 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Chris Sander and Reinhard Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, 9 (1):56–68, 1991. doi: 10.1002/prot.340090107.

Damiano Sgarbossa, Cyril Malbranke, and Anne-Florence Bitbol. ProtMamba: a homology-aware but alignment-free protein state space model. *Bioinformatics*, 41(6):btaf348, 2025. doi: 10.1093/bioinformatics/btaf348.

Dan S. Tawfik. Messy biology and the origins of evolutionary innovations. *Nature Chemical Biology*, 6(10):692–696, 2010. doi: 10.1038/nchembio.441.

Thomas C. Terwilliger, Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read, and Paul D. Adams. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods*, 21:110–116, 2024. doi: 10.1038/s41592-023-02087-4.

Jr. Truong, Timothy and Tristan Bepler. Understanding protein function with a multimodal retrieval-augmented foundation model. In *Advances in Neural Information Processing Systems*, volume 38, pp. 59170–59225, 2025. doi: 10.52202/085713-1772. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4d19160864e6a644496d61b21c7e015a-Abstract-Conference.html. PoET-2.

Jr. Truong, Timothy F. and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77379–77415, 2023. doi: 10.52202/075280-3384. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4366126eba252699b280e8f93c0ab2f-Abstract-Conference.html.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52309–52333. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24ct.html.

Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with FoldBench. *Nature Communications*, 17:442, 2026. doi: 10.1038/s41467-025-67127-3.

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, 2023. doi: 10.1126/science.adf2465.

Le Zhang, Jiayang Chen, Tao Shen, Yu Li, and Siqi Sun. MSA generation with seqs2seqs pretraining: Advancing protein structure predictions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 57324–57348, 2024. doi: 10.52202/079017-1827. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/694be3548697e9cc8999d45e8d16fe1e-Abstract-Conference.html.

Yang Zhang and Jeffrey Skolnick. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005. doi: 10.1093/nar/gki524.

A Implementation and Training Details

A.1 Model Architecture

Diffusion MSA Pairformer extends the MSA Pairformer architecture (Akiyama et al., 2026) with structured diffusion denoising capabilities. The model processes tokenized MSAs of dimension $N \times L$, where N is the number of sequences (MSA depth) and L is the alignment length.

Table 7: **DMP architecture configuration.**

Configuration	Value
Core blocks	22
Maximum MSA depth and aligned length	$N = 128,\ L = 512$
MSA and pair dimensions	$d_{\text{msa}} = 464,\ d_{\text{pair}} = 256$
MSA-update heads	8
Vocabulary size	28
Trainable parameters	111M

Each block updates the MSA representation by pair-weighted averaging and then updates the pair representation through a query-biased outer product followed by triangle multiplication. A continuous diffusion-time embedding and a shared mode embedding are added to the MSA representation before the backbone. The training configuration uses one shared mode id for all masking tasks; task identity is specified by the input corruption and loss masks rather than by task-specific embeddings.

A.2 Training Strategy

Optimization and objective configuration. Table 8 summarizes the optimization, diffusion, loss-weighting, and task-mixture parameters. The first epoch uses token-level diffusion as warm-up; subsequent epochs use the fixed task mixture. The column objective masks at most $C_{\max} = 15$ columns.

Table 8: **Training and optimization configuration.**

Parameter	Value
Learning rate	$4 \times 10^{-5} \to 10^{-5}$, linear decay
Warm-up	1,000 optimization steps
Training budget	20 epochs (134,340 optimization steps)
Diffusion schedule	$T = 500$, $\rho_t = t/T$, uniform timestep sampling
Task probabilities	$(\pi_{\text{token}}, \pi_{\text{row}}, \pi_{\text{col}}, \pi_{\text{query}}) = (0.30, 0.30, 0.25, 0.15)$
Column limit	$C_{\max} = 15$
Non-gap weights	homolog 1.0; query 2.0
Gap-weight warm-up	homolog $0.05 \to 0.20$; query $0.02 \to 0.10$
Gap-weight schedule	first 3,000 optimization steps

Time encoding. We encode $\rho_t = t/T$ using Fourier features with $R = 16$ frequency components:

$$\boldsymbol{\phi}(\rho_t) = [\rho_t, \rho_t^2, \sin(2\pi\rho_t), \cos(2\pi\rho_t), \ldots, \sin(2\pi R\rho_t), \cos(2\pi R\rho_t)]. \tag{5}$$

A two-layer MLP projects $\boldsymbol{\phi}(\rho_t)$ to the MSA embedding dimension, and the resulting embedding is added to every MSA position.

A.3 Training Data

Dataset. We start from an OpenProteinSet-derived UniClust30 accession index containing 268,778 query accessions (approximately 270k) (Ahdritz et al., 2023; Mirdita et al., 2017). After preprocessing, 215,356 MSAs satisfy the eligibility criteria; 214,925 are assigned to training and 431 to validation. The split contains no duplicated query identifiers or train–validation overlap. A separate sequence-identity audit found no training sequence with at least 95% identity to any protein in the evaluation tasks.

MSA preprocessing. Each protein serves as the query sequence of one A3M-format MSA. Before tokenization, each alignment is filtered with HHfilter using the recorded 95% identity, 50% coverage, and 15% query- identity settings:

```
hhfilter -i INPUT.a3m -o OUTPUT.a3m -id 95 -cov 50 -qid 15
```

Lowercase insertion characters are mapped to gaps and A3M gap symbols are normalized. Alignments with fewer than 32 sequences or a maximum gap ratio above 0.95 are excluded, and query sequences are deduplicated during corpus construction. The aligned length is cropped to at most $L = 512$.

Depth sampling. The query row is retained and homolog rows are sampled randomly during training, with total MSA depth capped at $N = 128$. Thus, each training example contains the query and at most 127 sampled homolog rows; the effective depth can be smaller when fewer homologs are available.

A.4 Decoding Algorithm

Algorithm 1 presents the iterative masked-diffusion decoding procedure used for inference. The algorithm progressively commits tokens from high-confidence to low-confidence positions using the implementation-aligned schedule in Equation 4.

Algorithm 1 Masked diffusion decoding for MSA generation

Require: Masked input MSA $\mathbf{X}$, denoising model p_θ, number of steps S
1: Initialize $\mathcal{U}_S \leftarrow$ all masked positions
2: **for** $r = S, S-1, \ldots, 1$ **do**
3: Set $t_r \leftarrow \max(1, \mathrm{round}(rT/S))$
4: Predict categorical distributions for all positions in $\mathcal{U}_r$ at t_r
5: Sample candidate tokens and compute confidence $\gamma_{ij}^{(r)}$ from model logits
6: $k_r \leftarrow$ commit size from Equation 4
7: $\mathcal{C}_r \leftarrow$ top-k_r positions in $\mathcal{U}_r$ by confidence $\gamma_{ij}^{(r)}$
8: Commit sampled tokens at positions in $\mathcal{C}_r$
9: $\mathcal{U}_{r-1} \leftarrow \mathcal{U}_r \setminus \mathcal{C}_r$
10: **end for**
11: **return** Completed MSA $\mathbf{X}$

The diffusion time and stepwise commit count are defined in Equation 4. The implementation clips each increment to the number of unresolved positions and commits all residual positions at the final reverse step.

Decoding hyperparameters. We use the following hyperparameters for iterative decoding across all generation tasks: temperature $\tau = 1.0$, number of steps $S = 8$. Sampling is performed over the full vocabulary at each masked position.

B Representation Benchmarks

B.1 Task Definitions and Evaluation Protocols

Representation inputs. Representation evaluations use the complete clean MSA as context and pass timestep $t = 0$ to the model, which clamps the time-conditioning input to its minimum valid

value, $t = 1$. For variant-effect scoring, only the scored query positions are additionally masked according to the masked-marginal protocol.

Variant-effect prediction. ProteinGym aggregates deep-mutational-scanning assays across diverse protein families and phenotypes, providing a standard benchmark for zero-shot mutation-effect prediction (Notin et al., 2023). Following the masked-marginal scoring convention (Meier et al., 2021), we score a substitution mutant x' that differs from the query sequence x at positions $\mathcal{M}$ by masking those positions in the aligned MSA context and contrasting the model's marginal likelihoods for the mutant and wild-type residues:

$$s(x' \to x) = \sum_{i \in \mathcal{M}} \Big[\log p_\theta\big(x'_i \mid x_{\setminus \mathcal{M}}\big) - \log p_\theta\big(x_i \mid x_{\setminus \mathcal{M}}\big) \Big], \tag{6}$$

where $x_{\setminus \mathcal{M}}$ denotes the query row with positions $\mathcal{M}$ masked and p_θ is the model's denoising marginal at each masked position. Within each assay of n variants, we rank the predicted scores $\{s_k\}$ against the experimental measurements $\{y_k\}$ and compute Spearman's rank correlation

$$\rho = 1 - \frac{6 \sum_{k=1}^{n} d_k^2}{n(n^2 - 1)}, \tag{7}$$

where d_k is the difference in ranks for variant k. We report the weighted mean of ρ across 217 assays covering 194 unique proteins, matching the ProteinGym leaderboard convention.

Intrachain contact prediction. Coevolving residue pairs are enriched among residues in physical contact within the folded structure, so recovering contacts from an MSA directly probes whether a model has captured this signal (Morcos et al., 2011; Rao et al., 2021). We evaluate on CASP15 targets under the contact-precision protocol used for MSA Pairformer (Akiyama et al., 2026), using its pretrained contact head on the pair representation from layer 15. The head is kept frozen. For a query protein of length L, let $\hat{P} \in [0,1]^{L \times L}$ be the pairwise contact probabilities and $C \in \{0,1\}^{L \times L}$ the ground-truth crystallographic contact map (Cβ–Cβ distance < 8 Å). Restricting to long-range pairs ($|i - j| \geq 24$) and taking the top-L predicted pairs $\mathrm{Top}_L(\hat{P})$,

$$P@L = \frac{|\mathrm{Top}_L(\hat{P}) \cap C|}{L}, \tag{8}$$

measures the fraction of top predictions that are true contacts. We report mean $P@L$ across targets.

Protein–protein interaction interface prediction. Coevolutionary constraints extend across interacting chains, so accurate interface-contact prediction requires evolutionary information beyond a single protein fold (Akiyama et al., 2026). We use the original pretrained MSA Pairformer ConFind head on the pair representation from layer 18 and keep it frozen. For a heteromeric complex with chains A and B, let $\hat{P}^{\mathrm{inter}} \in [0,1]^{L_A \times L_B}$ be the predicted inter-chain contact probabilities, C^{inter} the ground-truth interface contact map, and $K = \sum_{i,j} C^{\mathrm{inter}}_{ij}$ the number of true interface contacts. Interface precision at K is

$$P@K = \frac{|\mathrm{Top}_K(\hat{P}^{\mathrm{inter}}) \cap C^{\mathrm{inter}}|}{K}. \tag{9}$$

We evaluate on the 27-complex benchmark from Akiyama et al. (2026) and report median $P@K$ across complexes.

B.2 Case Studies

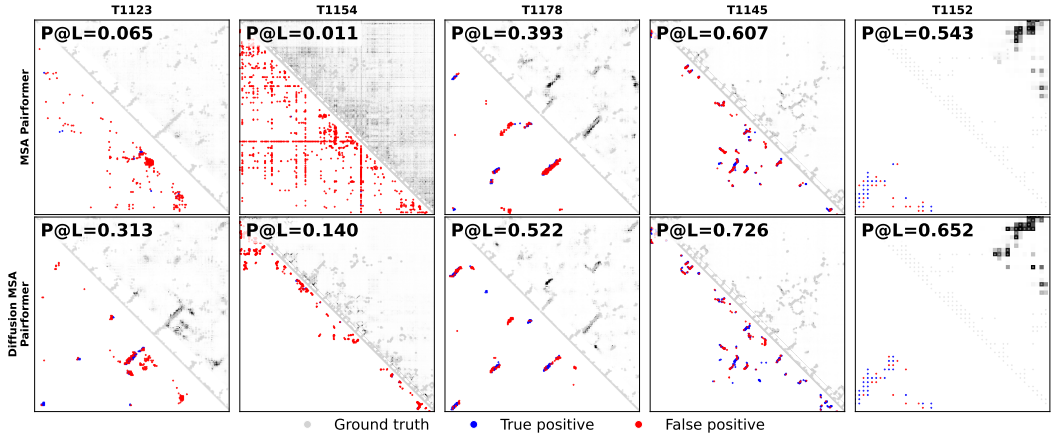

Figure 4: **CASP15 contact-map case studies.** Upper triangle: predicted contact probability. Lower triangle: top-L long-range predictions against crystallographic contact map (gray: ground truth, blue: true positive, red: false positive).

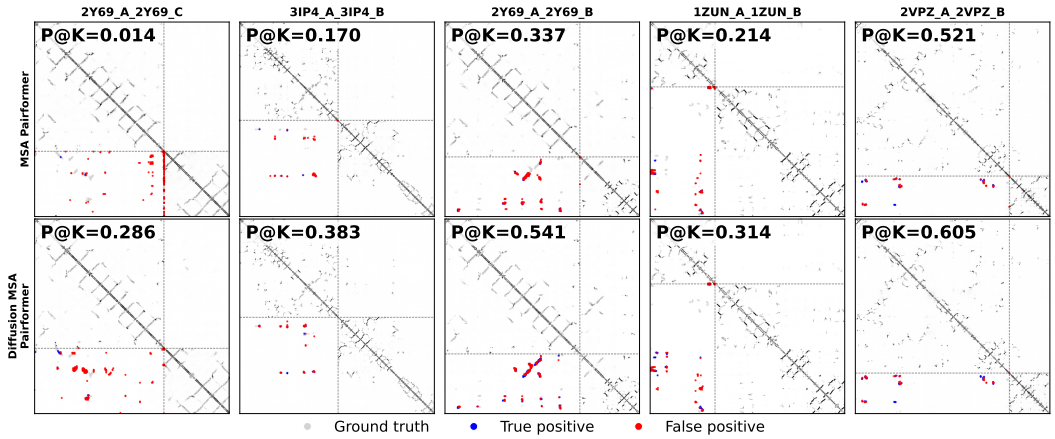

Figure 5: **PPI interface case studies.** Upper triangle: predicted inter-chain contact probability. Lower triangle: top-K predictions against crystallographic interface (gray: ground truth, blue: true positive, red: false positive). Dashed lines mark chain boundary.

B.3 Task ablation protocols and significance tests

All ablation variants use the same training setup except for the objective probabilities. The five configurations are: (1) the full structured mixture ($\pi_{\text{token}} = 0.30$, $\pi_{\text{row}} = 0.30$, $\pi_{\text{col}} = 0.25$, $\pi_{\text{query}} = 0.15$); (2) uniform token diffusion ($\pi_{\text{token}} = 1.0$); and (3–5) dropout-one-task variants in which the removed task's probability is reassigned to token diffusion. This design isolates task-level effects while holding the architecture and training pipeline constant.

For each ablation versus the full model, we test the one-sided null of no improvement using paired sign-flip permutation tests on per-case scores. The aggregate statistic for each benchmark—official weighted ρ for ProteinGym, mean per-complex $P@K$ for PPI-27, and mean per-target $P@L$ for CASP15—is recomputed after flipping the signs of the paired per-case differences. The p-value is the fraction of permuted statistics that exceed the observed statistic, using 10^5 permutations for PPI-27 and CASP15.

Benchmark	Ablation	Δ (full−abl.)	p (raw)	Stars
ProteinGym (ρ)	Uniform	$+0.0206$	5×10^{-5}	***
	w/o query	$+0.0009$	0.293	n.s.
	w/o row	$+0.0014$	0.178	n.s.
	w/o column	$+0.0070$	5×10^{-5}	***
PPI-27 ($P@K$)	Uniform	$+0.140$	10^{-5}	***
	w/o query	$+0.024$	0.025	*
	w/o row	$+0.023$	0.084	n.s.
	w/o column	$+0.032$	0.030	*
CASP15 ($P@L$)	Uniform	$+0.046$	10^{-5}	***
	w/o query	$+0.017$	0.0012	**
	w/o row	$+0.014$	0.0016	**
	w/o column	$+0.015$	0.022	*

Table 9: **Paired sign-flip permutation tests for each ablation against the full model.** Δ is the full-minus-ablation difference. For PPI-27 and CASP15, Δ is the mean paired difference over complexes or targets; for ProteinGym, it is the difference in the official weighted Spearman ρ. The reported p-values and significance stars are unadjusted.

C MSA Generation Details

C.1 Sampling Protocol

MSA depth selection. We generate MSAs with target depth 32, following the analysis in Chen et al. (2024), which demonstrated that depth 32 provides an effective balance between computational cost and structural prediction accuracy for orphan proteins. Depths below 32 may provide insufficient evolutionary signal, while depths substantially above 32 can inflate confidence scores without corresponding structural improvements (Akdel et al., 2022).

Generation modes. For each Hard52 case, we generate 40 candidate MSAs using two source modes:

- **Generate-to-32**: Use the available shallow MSA as context to generate replacement homolog rows, then retain the query and generated rows as a depth-32 MSA while discarding the original homolog rows.
- **Fill-to-32**: Preserve the available shallow MSA and generate only the additional rows required to reach depth 32.

Decoding strategies. Each source mode is combined with one-step sampling or iterative MaskGIT decoding:

- **One-step sampling**: At inference time, decode the model once at terminal diffusion time $t = T$, and sample generated tokens from all masked positions.
- **MaskGIT decoding**: Starting from the same fully masked row of MSA, repeatedly predict token probabilities, commit the highest-confidence unresolved positions, and retain the remaining positions as masks according to the implementation-aligned schedule in Equation 4.

We draw 10 independent samples from each of the four combinations (two generation modes $\times$ two decoding strategies), yielding 40 candidate MSAs per case.

C.2 Conditional-NLL Confidence Selector

Mean pLDDT alone can mis-rank candidate structures generated from MSAs. We therefore evaluate a DMP-specific selector that supplements downstream structure confidence with DMP's conditional likelihood for the added MSA rows. The conditional NLL is computed after generation and does not use native structures or TM-scores.

Formulation. Let R be the case's fixed raw shallow MSA, and let A_k be the manifest-defined added rows for candidate k. We keep R visible, replace all tokens in A_k with DMP's mask token,

and use one DMP forward pass to score the true added tokens. If Ω_k is the set of valid target-token positions in those rows, the normalized conditional NLL is

$$\ell_k = -\frac{1}{|\Omega_k|} \sum_{(r,j)\in\Omega_k} \log p_\theta(x_{k,rj} \mid R, \mathrm{mask}(A_k)), \tag{10}$$

where $x_{k,rj}$ is the observed token in the generated candidate. Averaging over valid tokens avoids making the score a simple function of alignment length or the number of added rows. Lower ℓ_k indicates that DMP assigns higher probability to the added rows given the fixed raw MSA. We combine this with mean pLDDT $\bar{p}_k$ using within-case z-scores:

$$s_k = z_{\mathrm{case}}(\bar{p}_k) - z_{\mathrm{case}}(\ell_k), \tag{11}$$

and select the candidate with the largest s_k. Thus higher pLDDT and lower conditional NLL both contribute positively. Standardization is computed within each case and the exact candidate pool for the corresponding structure predictor.

Algorithm. Algorithm 2 formalizes the ranking procedure. The scoring pass masks candidate-added rows while keeping the raw MSA visible; it does not alter the generated MSA or run an additional generation trajectory.

Algorithm 2 Conditional-NLL DMP selector

Require: DMP candidate MSAs $\{\mathbf{X}_k\}_{k=1}^{K}$, fixed raw MSA R, structure predictor f, DMP scoring model p_θ

for each candidate $\mathbf{X}_k$ **do**
  Predict structure $\hat{y}_k \leftarrow f(\mathbf{X}_k)$
  Compute mean pLDDT: $\bar{p}_k \leftarrow \mathrm{mean}(\mathrm{pLDDT}(\hat{y}_k))$
  Mask added rows A_k and compute mean conditional NLL ℓ_k using Eq. 10
  Compute selector score: $s_k \leftarrow z_{\mathrm{case}}(\bar{p}_k) - z_{\mathrm{case}}(\ell_k)$
end for
return $\arg\max_k s_k$

C.3 MSA Composition Analysis

MSA generation methods exhibit distinct sequence-composition and gap-distribution patterns. Figure 6 compares MSA visualizations for Hard52 case 4nwr_BG across ProfileBFN, MSAGPT, and Diffusion MSA Pairformer at depth 32.

Composition differences. MSAGPT generates substantially sparser MSAs with higher gap content and elevated sequence identity to the query. This sparse coverage correlates with MSAGPT's smaller gap between oracle TM-score and pLDDT top-1 TM-score in Table 1: fewer alignment tokens yield better-calibrated pLDDT scores. In contrast, ProfileBFN and Diffusion MSA Pairformer produce denser MSAs with higher amino acid coverage across aligned positions. While increased coverage provides richer evolutionary signal or more noise, it also elevates pLDDT values that do not always correlate with structural accuracy.

C.4 Structure Prediction Case Studies

Figures 7 and 8 present representative Hard52 predictions from the AlphaFold2 and ESMFold2 evaluations, respectively.

D Family-Conditioned Enzyme Design Details

D.1 Single-enzyme evaluation

The six families in Table 10 are sourced from the ProfileBFN and MSAFlow benchmarks (Gong et al., 2025; Anonymous, 2026). For P40925, Q7X7H9, and Q15165, every method generates 1000 sequences conditioned on the family MSA at temperature $T = 1.0$; for Q15BH7, P13280, and P57298, the comparison uses 100 sequences, matching the MSAFlow protocol. Each sequence is passed to CLEAN (Yu et al., 2023) for EC-label-set prediction, following the evaluation protocol used in ProfileBFN. In this mono-functional setting, a sequence is counted as accurate when its predicted EC label agrees with the family annotation. Acc. $\times$ Uniqueness is the accuracy rate multiplied by the fraction of generated sequences that are not exact duplicates, so a method cannot inflate its score by repeating a small number of correct sequences.

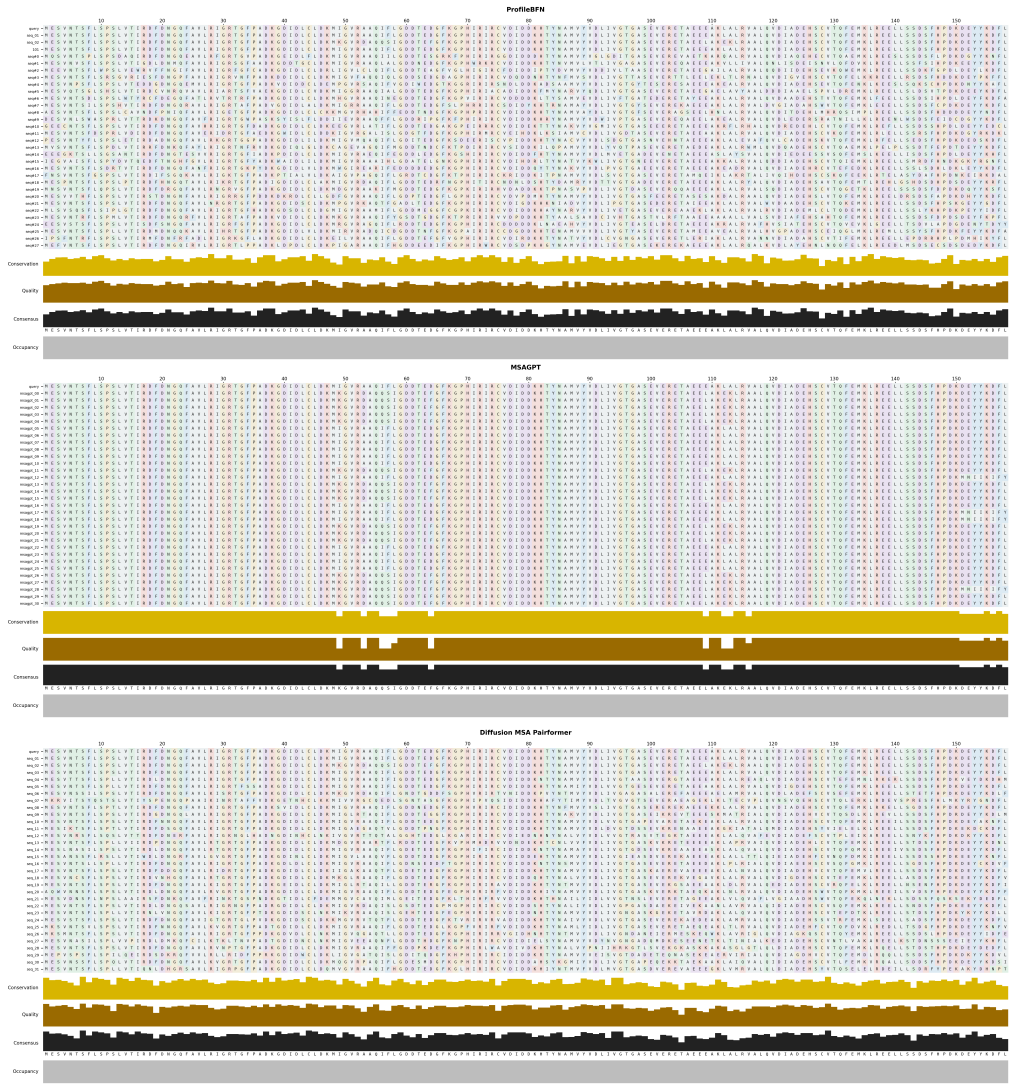

Figure 6: **MSA composition comparison (case 4nwr_BG).** ProfileBFN (top), MSAGPT (middle), and DMP (bottom), with the query row above homologs and gaps in white. The four tracks below each alignment show conservation, quality, consensus, and occupancy from top to bottom.

Table 10: **Single-enzyme design cases.**

ID	EC	Length	MSA Depth
P40925	1.1.1.37	334	572
Q7X7H9	2.7.1.71	287	443
Q15165	3.1.1.2	354	15
Q15BH7	2.3.1.82	152	12
P13280	2.4.1.186	333	13
P57298	3.6.1.55	124	15

D.2 MULTI-EC BENCHMARK CONSTRUCTION

The multi-EC benchmark is built from the CLEAN split100 dataset (Yu et al., 2023). We retain proteins annotated with at least two EC numbers at full four-level specificity and group them by their exact EC-label set, so every member of a family shares the same multi-functional annotation. Requiring 10–20 members and at most 90% pairwise identity leaves 15 eligible families. These families, listed in Table 11, span 219–537 residues and conditioning depths of 11–22.

For each family, every method generates 1000 sequences, and CLEAN (Yu et al., 2023) predicts a set of EC labels for each sequence. Let E_f be the reference EC set shared by family f and $\widehat{E}(x)$ the predicted set for generated sequence x. Any-EC accuracy is the fraction satisfying $\widehat{E}(x) \cap E_f \neq \varnothing$, whereas all-EC accuracy is the fraction satisfying $E_f \subseteq \widehat{E}(x)$. Pairwise identity is the mean sequence identity between generated sequence pairs, computed by global alignment; lower values indicate greater sequence diversity. Uniqueness is the fraction of generated sequences that are not exact duplicates.

Table 11: **Multi-EC design cases.**

ID	Multi-EC	Len	Enzyme	Depth
O28028	6.3.2.31; 6.3.2.34	249	Coenzyme F420:L-glutamate ligase	15
C3MJ44	3.5.1.130; 3.5.1.132	346	[LysW]-lysine/ornithine hydrolase	15
O35469	1.1.1.145; 5.3.3.1	352	3-oxo-5-beta-steroid 4-dehydrogenase	18
O52058	6.3.4.14; 6.4.1.2	537	Biotin carboxylase/ligase	14
O59397	1.2.1.103; 1.2.1.106	330	[LysW]-aminoadipate reductase	20
P0A9T0	1.1.1.399; 1.1.1.95	410	D-3-phosphoglycerate dehydrogenase	15
P0AA37	5.4.99.28; 5.4.99.29	219	RNA pseudouridine synthase RluA	13
P43636	2.4.1.132; 2.4.1.257	467	Glycosyltransferase	16
P80147	2.6.1.19; 2.6.1.22	500	4-aminobutyrate aminotransferase	11
P17618	1.1.1.193; 3.5.4.26	361	Riboflavin biosynthesis protein RibD	14
P51110	1.1.1.219; 1.1.1.234	337	Dihydroflavonol 4-reductase	13
P61800	3.6.1.71; 3.6.1.72	342	Aprataxin	19
A0QWT2	4.1.1.36; 6.3.2.5	414	CoA biosynthesis protein CoaBC	17
Q1M0V6	2.4.1.102; 2.4.1.148; 2.4.1.150	440	Beta-1,3-galactosyl-O-glycosyl transferase	13
Q47878	1.21.4.2; 1.21.4.3; 1.21.4.4	472	Anaerobic coproporphyrinogen-III oxidase	22

D.3 STRUCTURE-LIGAND EVALUATION

The six families in Table 12 are selected from the multi-EC benchmark because each has crystal structures resolved in multiple ligand-bound states. Across these families there are 16 reference states. Each method generates 1,000 sequences per family. We use ESMFold2 (Candido et al., 2026) for protein-only prediction and for ligand-aware co-folding, providing each generated sequence together with the CCD identifiers for the corresponding experimental state. This gives 16,000 co-folded complex predictions and 26,000 ligand–prediction pairs per method; states with multiple scored organic ligands contribute one pair per ligand. All 26,000 pairs per method received valid scores (100% score coverage). We evaluate all generated candidates without confidence-based ranking or best-of-N selection.

For protein-only predictions, pLDDT is ESMFold2's per-residue confidence and TM-score measures global fold similarity after alignment to the reference crystal with TM-align (Zhang & Skolnick, 2005). For co-folded predictions, global protein TM-score is computed separately from the local ligand and pocket scores, using the same protein-chain alignment protocol.

We use OpenStructure's SCRMSDScorer and LDDTPLIScorer (Biasini et al., 2013; Xu et al., 2026). SCRMSDScorer computes symmetry-corrected ligand RMSD (LRMSD) after binding-site superposition and returns lDDT-LP for the pocket; LDDTPLIScorer measures recovery of protein–ligand interatomic distances. A ligand–prediction pair is successful under the FoldBench criterion only if LRMSD is < 2 Å and lDDT-PLI is > 0.8.

Only the organic substrate, cofactor, or inhibitor CCD components listed for each state are included in the ligand-pose scores. Inorganic ions (MG, MN, ZN, CA), phosphate, iron–sulfur clusters, waters, and crystallization additives are excluded; metal-coordination geometry is therefore not evaluated. For each metric, we first average ligand–prediction pairs within each state, then average states equally within each family, and finally average the six families equally. The resulting family-macro values are descriptive; generated candidates are not independent biological replicates.

D.4 Full enzyme design cases

Table 12: **Full enzyme design cases.** Original: UniProt ID in CLEAN split100. Representative: ID used for structure evaluation (differs when structure availability requires alternative family member). Type: catalytic architecture. Depth: MSA depth. States: number of experimental ligand-binding/conformational states.

Rep.	Original	Multi-EC	Type	Len	Depth	States
O28028	A6VFM8	6.3.2.31; 6.3.2.34	Single pocket, multi-state	249	15	3
P0A9T0	P0A9T0	1.1.1.399; 1.1.1.95	Single pocket, multi-substrate	410	15	2
P80147	P50554	2.6.1.19; 2.6.1.22	Single pocket, multi-substrate	500	11	3
P17618	P50853	1.1.1.193; 3.5.4.26	Dual domain, dual pocket	361	14	3
P51110	P51109	1.1.1.219; 1.1.1.234	Single pocket, multi-substrate	337	13	2
A0QWT2	P67734	4.1.1.36; 6.3.2.5	Dual domain, dual pocket	414	17	3

Table 13: **Reference ligand states.** PDB structures and ligands for structure-ligand compatibility evaluation.

Family	States (PDB:Ligands)
O28028	7ULD: GTP+MN; 7ULE: F4I+GDP+MN; 8G8P: F42+GTP+GDP+MN
P0A9T0	1SC6: NAD; 1YBA: NAD+AKG+PO4
P80147	4Y0H: PLP+FES; 1OHW: PLP+VIG+FES; 4Y0I: PSZ+FES
P17618	3EX8: AIF+ZN; 2D5N: NDP+ZN; 4G3M: AOF+AI9+ZN
P51110	2NNL: NAP+ERD; 3BXX: NAP+QUE
A0QWT2	6TGV: CTP+FMN+MG; 4QJI: CTP+MG; 8OWR: CTP+W5E+CA

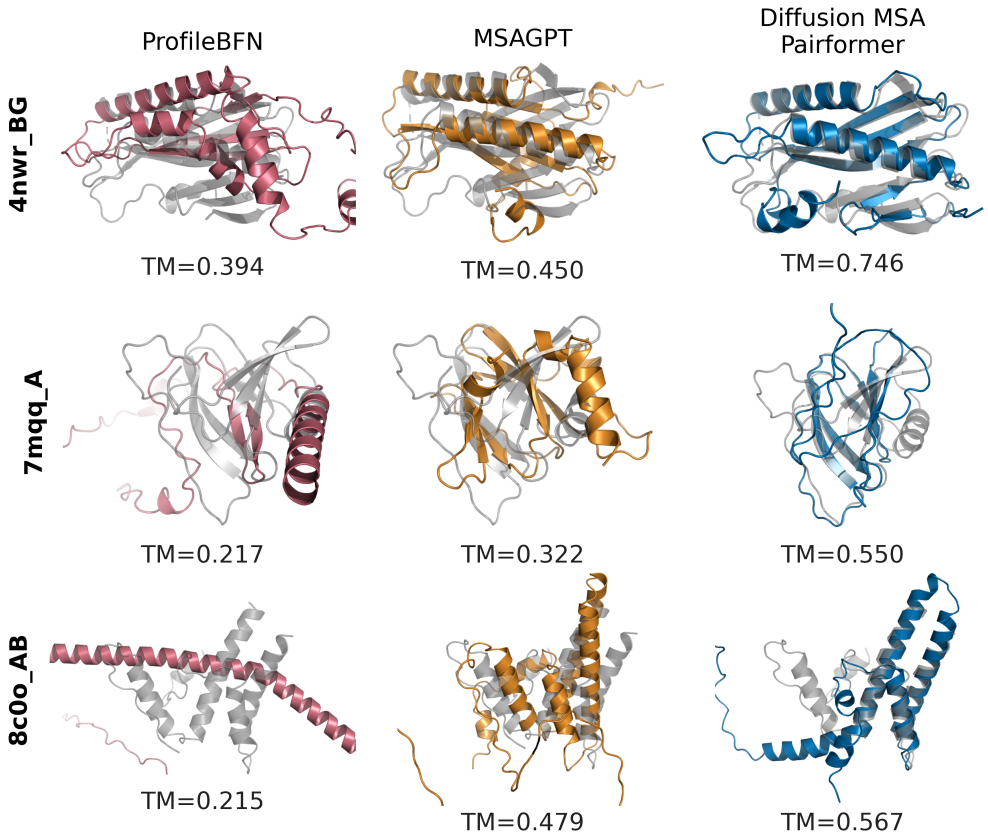

Figure 7: **AlphaFold2 case studies.** Three representative cases showing pLDDT top-1 predictions. Native structure overlaid in gray. TM-scores reported below predictions.

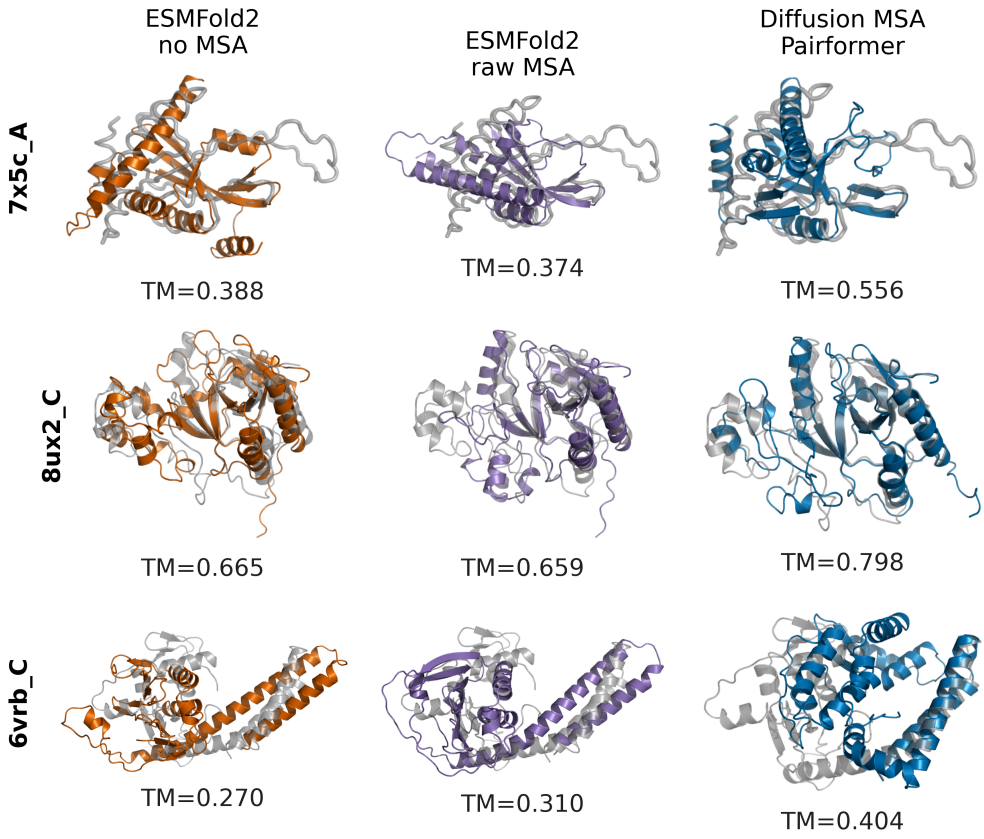

Figure 8: **ESMFold2 case studies.** Three representative cases comparing single-sequence, raw shallow MSA, and DMP-generated MSA predictions. Native structure overlaid in gray.