LATENTFOLD: CONTINUOUS MSA REPRESENTATIONS WITHOUT DISCRETE HOMOLOG GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple sequence alignments (MSAs) support accurate protein structure prediction by exposing conservation and coevolutionary signals. Existing MSA generative methods largely instantiate these signals as discrete homologous sequences, even though downstream folding models ultimately consume learned MSA representations. We ask whether useful MSA-derived information must take this discrete form. We propose **LatentFold**, a fully differentiable framework containing **LatentMSA**, a sequence-conditioned generator that instead represents structure-relevant MSA signals as a collection of continuous latent states. LatentMSA outputs an MSA-shaped latent tensor rather than discrete homologous sequences, producing distinct probe-conditioned states around a dynamic sequence-specific anchor. It is trained with a permutation-invariant statistical alignment objective that matches selected depth-aggregated moments and residue-pair similarity patterns of natural MSA representations, avoiding explicit correspondence between unordered homolog rows. Experiments on FoldBench and CASP15 show that LatentFold outperforms the evaluated discrete MSA-generation baselines across most concat-MSA-depth comparisons, with the clearest shallow-depth gains on FB_PROTEIN and more metric-dependent improvements on CASP15, while requiring no target-specific homology retrieval at inference time. Its differentiable design allows structural-loss gradients to pass through the frozen folding pipeline and update LatentMSA while the PLM and folding backbone remain frozen. These results support continuous latent MSA generation as a trainable alternative to discrete homolog generation for supplying structure-relevant MSA signals. The code is available at: `https://anonymous.4open.science/r/LatentFold-2C68`.

1 INTRODUCTION

Accurate three-dimensional protein structures are fundamental to understanding biological function and advancing structure-based drug design (Anderson, 2003; Renaud et al., 2018). Recent deep learning models, exemplified by AlphaFold2 and AlphaFold3 (Jumper et al., 2021; Abramson et al., 2024), have achieved near-atomic accuracy by extracting coevolutionary constraints from multiple sequence alignments (MSAs) through attention mechanisms. However, their performance remains highly dependent on retrieved MSA depth and quality, constituting a major vulnerability of retrieval-dependent architectures. For orphan proteins, low-homology families, and rapidly evolving pathogen targets, insufficient homologous information in public databases limits recovery of reliable covariance patterns and can substantially degrade prediction accuracy (Zhang et al., 2023).

To mitigate this homology bottleneck, recent efforts have explored protein language models and MSA generative approaches (Lin et al., 2023; Ferruz et al., 2022). MSA-free PLMs attempt to bypass explicit homology retrieval by learning evolutionary regularities from large-scale protein sequences, whereas MSA generative methods seek to augment poor alignments by synthesizing homologous sequences. Representative approaches include MSA Transformer-based mask-and-reveal generation (Rao et al., 2021), EvoGen (Zhang et al., 2023), MSA-Generator (Zhang et al., 2024a), MSAGPT (Chen et al., 2024), and PLAME (Cao et al., 2025). These methods have improved structure prediction in low-homology regimes by enriching MSA-like inputs. However, they inherit an implicit representational assumption: that useful MSA-derived signals must first be instantiated as discrete homolog-like sequences. This sequence-to-sequence generative paradigm faces several practical limitations. *First*, exploring the discrete amino-acid space is expensive, especially when

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.

generating deep synthetic MSAs with many homolog-like rows. *Second*, token-level generation may produce sequence-plausible homologs without ensuring consistent coevolutionary signals across generated MSA rows, especially for long-range residue pairs relevant to folding; generated homologs may therefore provide noisy or ambiguous signals to downstream predictors. *Third*, sampling or argmax operations over amino-acid identities hinder direct gradient propagation from downstream folding objectives to the MSA generator, typically requiring straight-through estimators, continuous relaxations, or high-variance policy-gradient methods.

More fundamentally, non-query MSA rows have permutation symmetry: reordering homologs does not change the conservation and covariation statistics considered by structure predictors (Jumper et al., 2021). Token- or sequence-level reconstruction objectives in discrete sequence space (Zhang et al., 2023; Chen et al., 2024; Zhang et al., 2024a; Cao et al., 2025) do not enforce this symmetry and can introduce unnecessary row correspondence. Explicitly resolving such correspondence would require a set-matching or assignment procedure. Together, these observations suggest that discrete homologs are a conventional carrier of evolutionary information, but need not be the representation supplied to an MSA-based folding model. They instead motivate a continuous objective that preserves structure-relevant MSA signals, is invariant to homolog-row permutations under the selected statistics, and remains differentiable with downstream folding modules.

To tackle these challenges, we propose **LatentFold**, a fully differentiable protein structure prediction framework that performs no target-specific homology retrieval at inference time and contains a **Dynamic-Anchor State-Conditioned Latent MSA Generator**. Rather than generating discrete homologous sequences, its LatentMSA module represents structure-relevant MSA signals as a sequence-conditioned collection of continuous latent states. It constructs an MSA-shaped representation directly in continuous latent space, deriving a sequence-conditioned anchor from single-sequence features and emitting N distinct probes through scale–direction decoupled perturbations. State-specific self-attention then produces probe-conditioned residue-dependency representations (Marks et al., 2011; Morcos et al., 2011). To avoid explicit homolog-row correspondence, LatentMSA is trained with a permutation-invariant objective inspired by Bardes et al. (2022). The objective aligns selected depth-aggregated moments and residue-pair similarity patterns between the generated representation and natural-MSA latent features; it does not assume recovery of particular homologs or the complete natural-MSA distribution.

LatentFold is modular and fully differentiable, supporting distinct supervision regimes. In the offline regime, LatentMSA maps single-sequence protein language model representations to MSA-shaped latent states by matching permutation-invariant statistics of natural MSA representations, without target-specific homology retrieval at inference time. When structural supervision is available, gradients from an AF3-style folding objective can pass through the folding pipeline to adapt LatentMSA (Jumper et al., 2021; Abramson et al., 2024; Team et al., 2025; Wohlwend et al., 2024; Chai Discovery, 2024). In our experiments, the PLM and folding backbone remain frozen. Our contributions are three-fold: **(i)** we formulate and empirically evaluate continuous latent MSA generation as an alternative to instantiating structure-relevant MSA signals through discrete homolog sampling, while avoiding target-specific homology retrieval at inference time and explicit row-wise correspondence; **(ii)** we introduce a dynamic-anchor, state-conditioned latent MSA generator trained by aligning selected permutation-invariant moments and residue-pair similarity statistics with natural-MSA latent representations; and **(iii)** we evaluate LatentFold under both offline natural-MSA-supervised statistical matching without structural labels and structure-supervised LatentMSA adaptation through a frozen folding backbone, observing gains across most concat-MSA-depth comparisons, with the clearest shallow-depth improvements on FB_PROTEIN and more metric-dependent gains on CASP15. Together, these contributions support the view that useful MSA-derived signals need not be instantiated as discrete homologous sequences, but can instead be supplied to folding models through a trainable collection of continuous latent states.

2 Related Work

Discrete Generative Models for MSAs. Recent protein language models have advanced MSA modeling and generation as a promising strategy for alleviating low-homology bottlenecks in protein structure prediction. MSA Transformer introduced an MSA-aware masked language modeling framework with row and column attention, providing a foundation for mask-based MSA infilling

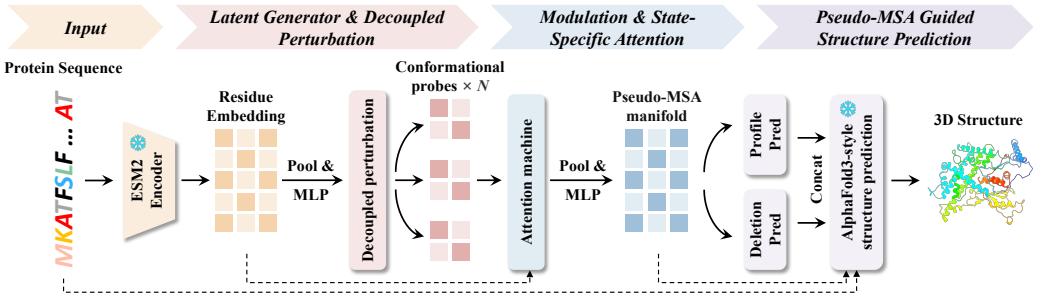


Figure 1: Conceptual overview of LatentFold. Rather than discrete homologous sequences, LatentMSA generates continuous probe-conditioned latent states trained by permutation-invariant alignment to selected natural-MSA statistics. No target-specific retrieval is performed at inference time. During structure-supervised adaptation, structural-loss gradients pass through frozen Protenix and update LatentMSA only; full details appear in Appendix A.6.

and structure-relevant representation learning (Rao et al., 2021). Subsequent methods, including EvoGen (Zhang et al., 2023), MSA-Generator (Zhang et al., 2024a), MSAGPT (Chen et al., 2024), and PLAME (Cao et al., 2025), synthesize or augment pseudo-MSAs from single sequences or sparse homologous inputs to enhance downstream folding. However, these approaches instantiate MSA-derived signals as discrete homolog-like sequences and use token- or sequence-level objectives that need not enforce permutation invariance over non-query rows or direct structural-gradient propagation. LatentFold instead represents selected structure-relevant MSA-derived signals continuously.

Moment Matching and Distribution Alignment. Moment matching and distribution alignment train generative models without explicit likelihoods or adversarial objectives. Maximum Mean Discrepancy (MMD) matches distributions by minimizing distances between kernel mean embeddings in a reproducing kernel Hilbert space (Gretton et al., 2012), and has been used in Generative Moment Matching Networks (GMMNs) for likelihood-free generation (Li et al., 2015; Dziugaite et al., 2015). In representation learning and domain adaptation, Deep CORAL aligns second-order feature statistics to reduce distribution shifts (Sun & Saenko, 2016), while variance–covariance regularization methods demonstrate the utility of decoupling feature variance and correlation constraints (Bardes et al., 2022). Inspired by these ideas, LatentFold uses natural MSAs as statistical supervision, training its latent representation through permutation-invariant alignment of selected depth-aggregated moments and residue-pair similarities rather than row-wise sequence reconstruction.

3 LatentFold

3.1 Problem Formulation and Preliminaries

Let $s = (s_1, s_2, \ldots, s_L)$ denote a protein sequence of length L. Single-sequence PLMs (e.g., ESM-2) map s to a contextual residue representation $\mathbf{H}_{\text{esm}} \in \mathbb{R}^{L \times d_{\text{esm}}}$. Although $\mathbf{H}_{\text{esm}}$ captures rich sequence-contextual information, it does not explicitly provide the residue–residue coupling statistics available from MSAs, which are important for accurate structure prediction. We nevertheless observe a weak association between ESM-2 feature similarities and MSA-derived pairwise patterns; Appendix A.1 provides the analysis and visualization.

Retrieval-based methods construct MSAs from large sequence databases. For each target, the retrieved MSA has a variable depth N_t and is embedded by a frozen MSA projection into a natural-MSA reference representation $\mathbf{M}_{\text{true}} \in \mathbb{R}^{N_t \times L \times c_m}$, where c_m is the channel dimension. Because homolog rows have no prescribed correspondence and N_t varies across proteins, fixed row-wise reconstruction is ill-defined. As illustrated in Appendix Figure 4(b), the selected row-level statistics are aggregated symmetrically across MSA depth, avoiding explicit homolog-row correspondence.

Terminology. We use **LatentMSA** to denote the sequence-conditioned generator $\mathcal{F}_\theta$. Its output $\mathbf{M}_{\text{pseudo}} \in \mathbb{R}^{N \times L \times c_m}$ is called a **pseudo-MSA**: it is an MSA-shaped continuous latent tensor, not an alignment of generated amino-acid sequences. **LatentFold** denotes the complete structure-prediction framework comprising a frozen single-sequence PLM, LatentMSA and its auxiliary MSA-feature

heads, and the downstream folding model. Natural MSAs provide training supervision, but no target-specific homology retrieval is performed at inference time. The goal is to preserve selected structure-relevant MSA-derived signals, not to recover individual homologs or the complete natural-MSA distribution.

Problem Statement: Given s and $\mathbf{H}_{\mathrm{esm}}$, we learn $\mathcal{F}_\theta$ to produce a fixed-depth pseudo-MSA representation:

$$\mathbf{M}_{\mathrm{pseudo}} = \mathcal{F}_\theta(\mathbf{H}_{\mathrm{esm}}) \in \mathbb{R}^{N \times L \times c_m}, \tag{1}$$

where N is the generated probe depth. Rather than autoregressively sampling discrete amino-acid sequences or imposing row-wise reconstruction, LatentMSA trains the generated latent representation with permutation-invariant statistical alignment, matching selected moments and residue-pair similarity statistics of $\mathbf{M}_{\mathrm{pseudo}}$ with those of $\mathbf{M}_{\mathrm{true}}$.

3.2 The LatentFold Architecture

LatentFold is a fully differentiable framework that performs no target-specific homology retrieval at inference time. As shown in Figure 1, LatentMSA maps ESM-2 embeddings to continuous pseudo-MSA representations whose selected statistics are aligned with natural-MSA latent representations. The pseudo-MSA and query sequence are then supplied to an AlphaFold-style architecture. LatentFold comprises three components: (1) Latent Generator & Decoupled Perturbation, (2) Modulation & State-Specific Attention, and (3) Decoupled Manifold Alignment Loss. The full data flow appears in Appendix Figure 6.

3.2.1 Latent Generator & Decoupled Perturbation

Given the ESM-2 representation $\mathbf{H}_{\text{esm}} \in \mathbb{R}^{L \times d_{\text{esm}}}$, we apply global pooling followed by an MLP projection to obtain a sequence-conditioned anchor $\mathbf{e}_{\text{anchor}} \in \mathbb{R}^{d_q}$, which serves as the center of the generated probe set. Appendix A.5 presents a qualitative folding-landscape analogy without assigning thermodynamic meaning to the anchor. We introduce an orthogonally initialized learnable direction matrix $\mathbf{P} \in \mathbb{R}^{N \times d_q}$, where each row defines a distinct latent direction and N is the pseudo-MSA depth. The probes $\mathbf{E}_{\text{probes}} \in \mathbb{R}^{N \times d_q}$ are formed by perturbing the anchor along normalized directions, with the learnable scalar λ_{radius} controlling the shared scale:

$$\mathbf{E}^{(n)}_{\text{probes}} = \mathbf{e}_{\text{anchor}} + \lambda_{\text{radius}} \cdot \frac{\mathbf{P}^{(n)}}{\|\mathbf{P}^{(n)}\|_2}. \tag{2}$$

3.2.2 Modulation & State-Specific Attention

To construct diverse latent states encoding residue-dependency patterns, we introduce a probe-conditioned gating mechanism that selectively modulates single-sequence representations. The projection weights operate on different sources: $\mathbf{W}^{\text{gate}}_q$ projects the latent probes $\mathbf{E}_{\text{probes}} \in \mathbb{R}^{N \times d_q}$, while $\mathbf{W}^{\text{gate}}_k$ and $\mathbf{W}^{\text{gate}}_v$ project the ESM-2 residue representations $\mathbf{H}_{\text{esm}} \in \mathbb{R}^{L \times d_{\text{esm}}}$. In a multi-head formulation, this yields $\mathbf{Q}_{\text{gate}} \in \mathbb{R}^{H \times N \times d_h}$ and $\mathbf{K}_{\text{gate}}, \mathbf{V}_{\text{gate}} \in \mathbb{R}^{H \times L \times d_h}$, where H is the number of heads and $d_h = d_q/H$.

A state-residue gating mask $\mathbf{A}_{\text{mask}} \in \mathbb{R}^{H \times N \times L}$ is computed as:

$$\mathbf{A}_{\text{mask}} = \sigma\left(\mathrm{Norm}_{\mathrm{gate}}\left(\frac{\mathbf{Q}_{\text{gate}}\mathbf{K}^{\top}_{\text{gate}}}{\sqrt{d_h}}\right)\right), \tag{3}$$

where $\mathrm{Norm}_{\mathrm{gate}}$ denotes the gate-logit normalization used before sigmoid activation; it is distinct from the row-wise L_2 normalization used in the alignment loss. This normalization stabilizes the logits and mitigates saturation Ramapuram et al. (2025). Each row of $\mathbf{A}_{\text{mask}}$ specifies how a latent probe attends to residues under a particular probe-conditioned latent state Qiu et al. (2025). The value features are then modulated by this mask and concatenated across heads to obtain the state-conditioned tensor:

$$\mathbf{V}_{\text{mod}} = \mathrm{Concat}^{H}_{h=1}\left(\mathbf{A}^{(h)}_{\text{mask}} \odot \mathbf{V}^{(h)}_{\text{gate}}\right) \in \mathbb{R}^{N \times L \times d_q}. \tag{4}$$

Finally, state-specific self-attention is applied independently along the residue dimension of each latent probe, enabling each generated pseudo-MSA row to capture long-range residue dependencies conditioned on its own probe-conditioned latent state.

State-Specific Self-Attention (SSSA). The gating module performs probe-conditioned residue modulation. To model non-local residue dependencies, we feed $\mathbf{V}_{\text{mod}}$ into the SSSA module. Queries, keys, and values are projected from the modulated features and split into H heads of width d_h:

$$\mathbf{Q}_{\text{self}} = \mathbf{V}_{\text{mod}}\mathbf{W}_q^{\text{self}}, \quad \mathbf{K}_{\text{self}} = \mathbf{V}_{\text{mod}}\mathbf{W}_k^{\text{self}}, \quad \mathbf{V}_{\text{self}} = \mathbf{V}_{\text{mod}}\mathbf{W}_v^{\text{self}}.$$

After head splitting, $\mathbf{Q}_{\text{self}}^{(h,n)}, \mathbf{K}_{\text{self}}^{(h,n)}, \mathbf{V}_{\text{self}}^{(h,n)} \in \mathbb{R}^{L\times d_h}$. Self-attention is computed independently for each head and latent probe along the residue dimension:

$$\mathbf{M}_{\text{attn}}^{(h,n)} = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{self}}^{(h,n)}\mathbf{K}_{\text{self}}^{(h,n)\top}}{\sqrt{d_h}}\right)\mathbf{V}_{\text{self}}^{(h,n)}, \quad h = 1,\dots,H,\ n = 1,\dots,N. \tag{5}$$

The resulting probe-conditioned $L \times L$ attention maps provide residue–residue dependency representations. Because each probe is modulated by a distinct latent direction, SSSA produces N probe-specific latent states rather than a single consensus representation. The head outputs are concatenated, normalized, and projected to the target channel dimension, producing the pseudo-MSA tensor $\mathbf{M}_{\text{pseudo}} \in \mathbb{R}^{N\times L\times c_m}$.

3.2.3 Decoupled Manifold Alignment Loss

A key challenge in latent MSA generation is the permutation symmetry of the non-query depth dimension under the statistics considered here. Homologous rows do not have a prescribed correspondence, and their number N_t varies across targets. Enforcing row-wise reconstruction would therefore introduce an unnecessary assignment problem and require matching representations with different depths. LatentFold instead aligns selected statistics aggregated symmetrically across depth. Here, "manifold alignment" denotes this statistical matching objective and does not imply recovery of the complete geometry or distribution of natural-MSA latent representations.

Given the generated pseudo-MSA $\mathbf{M}_{\text{pseudo}} \in \mathbb{R}^{N\times L\times c_m}$ and natural-MSA latent representation $\mathbf{M}_{\text{true}} \in \mathbb{R}^{N_t\times L\times c_m}$ from a frozen MSA projection module, we compute row-level statistics $\mathcal{S}(\mathbf{m}_n) = \{\mu_n, \sigma_n, \mathbf{C}_n, \mathbf{R}_n\}$ over residue positions, where $\mathbf{R}_n$ is the residue-pair similarity matrix for row n. For a representation with depth D, define $\bar{\mathcal{S}}(\mathbf{M}) = D^{-1}\sum_{n=1}^{D}\mathcal{S}(\mathbf{m}_n)$. For any homolog-row permutation π, we have:

$$\bar{\mathcal{S}}(\pi(\mathbf{M})) = \frac{1}{D}\sum_{n=1}^{D}\mathcal{S}(\mathbf{m}_{\pi(n)}) = \bar{\mathcal{S}}(\mathbf{M}), \tag{6}$$

making supervision independent of row order. Inspired by moment matching and variance–covariance regularization (Gretton et al., 2012; Li et al., 2015; Dziugaite et al., 2015; Sun & Saenko, 2016; Bardes et al., 2022), these aggregated statistics are matched between $\mathbf{M}_{\text{pseudo}}$ and $\mathbf{M}_{\text{true}}$:

$$\mathcal{L}_{\text{style}} = \mathcal{L}_{\text{mean}} + \lambda_{\text{std}}\mathcal{L}_{\text{std}} + \lambda_{\text{corr}}\left(\mathcal{L}_{\text{corr}}^{C} + \alpha_{\text{spatial}}\mathcal{L}_{\text{corr}}^{L}\right). \tag{7}$$

Here, $\mathcal{L}_{\text{mean}}$ aligns the depth-averaged channel mean, while $\mathcal{L}_{\text{std}}$ preserves within-row variation over residue positions. The two correlation terms capture complementary information: $\mathcal{L}_{\text{corr}}^{C}$ matches depth-averaged channel dependencies, whereas $\mathcal{L}_{\text{corr}}^{L}$ aligns depth-averaged residue-pair similarities over sequence-separated pairs. In particular, $\mathbf{C} \in \mathbb{R}^{c_m\times c_m}$ constrains feature-channel dependencies, while $\mathbf{R} = \mathbb{E}_n[\mathbf{R}_n] \in \mathbb{R}^{L\times L}$ is used as a representation-derived coupling surrogate. Because all terms are computed after symmetric aggregation over depth, the objective is invariant to homolog-row permutations under these statistics and does not require $N = N_t$.

Statistical alignment may still allow redundant latent probes: multiple probes can satisfy the selected aggregate constraints while producing similar representations. We therefore introduce a margin-based

probe diversity regularizer motivated by metric-learning margin enforcement (Guzman-Rivera et al., 2012; Schroff et al., 2015):

$$\mathcal{L}_{\text{div}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \max\left(0, \cos(\mathbf{E}_i, \mathbf{E}_j) - \tau_{\text{max}}\right), \tag{8}$$

where $\mathbf{E}_i \equiv \mathbf{E}_{\text{probes}}^{(i)}$ and τ_{max} is a similarity margin that discourages redundant probes while allowing related latent states. This regularizer encourages distinct probe embeddings around the dynamic anchor, complementing the depth-aggregated statistical objective. The final objective combines manifold alignment, probe diversity, and auxiliary statistical feature losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{style}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}} + \mathcal{L}_{\text{aux}}. \tag{9}$$

Detailed definitions of the individual moment terms are provided in Appendix A.7, and the auxiliary profile and deletion-mean losses are described in Appendix A.8.

3.2.4 Downstream Integration and End-to-End Differentiability

To integrate the pseudo-MSA with AlphaFold3-style structure modeling, LatentFold introduces two auxiliary heads: MSA profile and deletion mean. The pseudo-MSA tensor $\mathbf{M}_{\text{pseudo}}$ is average-pooled along the probe dimension and passed through a shared MLP to produce residue-level statistics. To stabilize optimization, we use task-specific auxiliary losses: KL divergence (Kim et al., 2021) for the probabilistic profile and log-smoothed Smooth L1 loss (Girshick, 2015) for the long-tailed deletion mean, mitigating gradient instability. The predicted profile, deletion mean, and generated latent MSA remain differentiable within the complete folding pipeline. In the experiments, structural-loss gradients pass through the frozen structure predictor to update LatentMSA only; both the PLM and folding backbone remain frozen. The same interface permits future joint fine-tuning of the complete pipeline. Further details are provided in Appendix A.8.

4 Empirical Validation

4.1 Datasets

For training, we used a Protein Data Bank (PDB) dataset (Burley et al., 2023) with a release cutoff of September 30, 2021. This dataset is a subset of the training corpus used by our backbone, Protenix (Team et al., 2025). We retained sequences of 32–1,000 residues and constructed MSAs using MMseqs2 (Steinegger & Söding, 2017; Mirdita et al., 2022). Homologs with gap fractions above 25% were removed, and each MSA was capped at 512 sequences. We augmented the data by using the top-k ($k = 3$) gapless homologs in each MSA as alternative queries, yielding 247,093 sequence–MSA pairs.

For evaluation, we used the monomeric FoldBench dataset (Xu et al., 2025), comprising 334 PDB targets (FB_PROTEIN) released between January 3, 2023, and November 1, 2024. To prevent leakage, all targets were filtered to have at most 40% sequence identity to the training set. We prioritize FoldBench over CAMEO (Haas et al., 2018) because it offers broader coverage and contains more targets in shallow retrieved-MSA regimes. CASP15 was included as a supplementary benchmark for challenging cases. For post-hoc stratification, we define each target's concat MSA depth as the sum of its UniRef100 (The UniProt Consortium, 2023) and ColabFoldDB (Mirdita et al., 2022) MSA depths, $d_{\text{concat}} = d_{\text{UniRef100}} + d_{\text{ColabFoldDB}}$, and group targets by d_{concat}. These retrieved MSAs are used only to assign evaluation strata and are not provided to LatentMSA, EvoGen, EvoDiff, MSAGPT, or Protenix NO MSA during inference. Detailed statistics are reported in Appendix Table 4.

4.2 Baselines

We use EvoDiff (Alamdari et al., 2023), EvoGen (Zhang et al., 2023), and MSAGPT (Chen et al., 2024) as MSA-generation baselines. Offline, all MSA-generation methods receive only the query sequence and perform no target-specific homology retrieval at inference time. We use Protenix, a structure predictor based on AlphaFold 3, as the shared downstream folding backbone; Protenix with only the query sequence ("Protenix NO MSA") serves as a no-MSA control. We also include

Table 1: Coupling agreement with $\mathbf{M}_{\text{true}}$ on FB_PROTEIN targets. Values are means across targets; higher is better. Best and second-best values are shown in **bold** and underlined, respectively. Metrics are defined in Appendix A.4.

Method	Spearman (all) ↑	Spearman (LR) ↑	Ref. overlap@$L/2$ ↑	Ref. overlap@L ↑	Ref. overlap@$2L$ ↑	q_{eff} ↑	ECF ↑
EvoDiff	0.080	0.064	0.011	0.021	0.037	0.001	0.001
EvoGen	0.288	0.277	0.127	0.180	0.251	0.050	0.053
MSAGPT	0.547	0.539	0.166	0.213	0.267	0.115	0.130
LatentMSA$_{\text{ESM2-8M}}$	0.478	0.466	0.283	0.309	0.343	0.144	0.168
LatentMSA$_{\text{ESM2-35M}}$	0.516	0.504	0.364	0.390	0.417	0.197	0.245
LatentMSA$_{\text{ESM2-150M}}$	0.607	0.598	0.431	0.454	0.482	0.272	0.373
LatentMSA$_{\text{ESM2-650M}}$	0.641	0.632	0.477	0.502	0.522	0.317	0.464
LatentMSA$_{\text{ESM2-3B}}$	**0.646**	**0.637**	0.496	0.523	0.542	**0.333**	**0.499**
LatentMSA$_{\text{ESM2-15B}}$	0.580	0.569	**0.513**	**0.533**	**0.543**	0.303	0.434

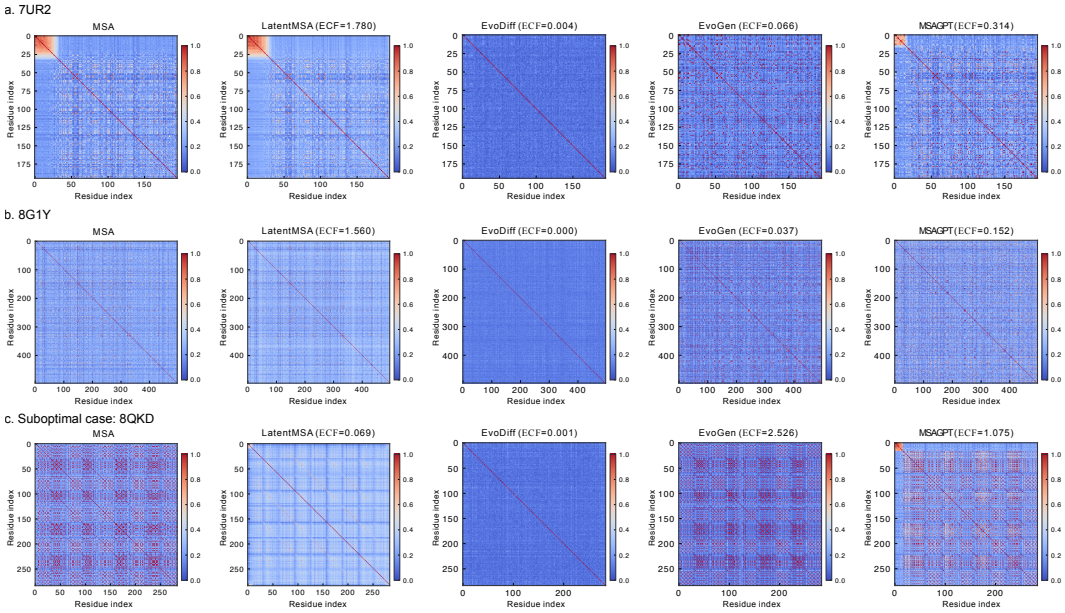


Figure 2: Representative coupling-surrogate matrices for LatentMSA, discrete MSA generators, and natural-MSA reference. ECF values follow the order LatentMSA/EvoDiff/EvoGen/MSAGPT. LatentMSA leads on **(a) 7UR2** ($1.780/0.004/0.066/0.314$) and **(b) 8G1Y** ($1.560/0.000/0.037/0.152$); EvoGen leads on **(c) 8QKD** ($0.069/0.001/2.526/1.075$).

ESMFold and SimpleFold (Wang et al., 2026) as strong single-sequence upper-bound references rather than directly comparable baselines, because their end-to-end structure-supervised training differs from the shared-backbone MSA-generation setting.

For structure evaluation, every model receives the same $K = 25$ budget: five seeds and five samples per seed for each target and method. We report metric-specific oracle scores by selecting the best of these predictions. This matched-budget protocol estimates attainable performance rather than single-sample or confidence-selected performance. Appendix A.13 additionally reports ranking-score Top-1 results on FB_PROTEIN and CASP15.

4.3 Quantitative Evaluation of Coevolutionary Signal Fidelity

The natural MSA and discrete MSAs from EvoDiff, EvoGen, and MSAGPT are mapped through the same frozen MSA projection into a structure-informed continuous space; LatentMSA predicts representations in this space. We compute identical negative pairwise L_2-distance coupling surrogates for all sources and evaluate rank agreement, reference-coupling overlap, and *Effective Coupling Fidelity* (ECF). Definitions and limitations are provided in Appendices A.3 and A.4.

Table 2: Per-target best-of-25 performance on FB_PROTEIN, stratified by concat MSA depth, $d_{\mathrm{UniRef100}} + d_{\mathrm{ColabFoldDB}}$. Values are mean ± SD within each bucket; bold and underline mark the best and second-best results. LatentMSA also has substantially lower measured input-construction time than the timed retrieval and generation methods; see Appendix A.10.

Method	[1,100), N=33				[100,500), N=47				[500,16k), N=254			
	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑
ESMFold	<u>0.694</u> ±0.206	5.301 ±5.518	<u>0.748</u> ±0.247	<u>0.732</u> ±0.258	0.704 ±0.197	5.110 ±5.168	<u>0.784</u> ±0.229	<u>0.730</u> ±0.236	<u>0.836</u> ±0.097	2.960 ±3.661	<u>0.911</u> ±0.115	**0.867** ±0.132
SimpleFold	0.653 ±0.197	5.802 ±5.236	0.722 ±0.241	0.691 ±0.260	0.669 ±0.160	<u>4.990</u> ±4.532	0.769 ±0.207	0.700 ±0.211	0.805 ±0.091	<u>2.729</u> ±2.640	0.904 ±0.102	0.850 ±0.123
PX(NO MSA)	0.676 ±0.239	5.284 ±5.784	0.712 ±0.274	0.688 ±0.306	0.531 ±0.205	11.609 ±5.659	0.465 ±0.226	0.391 ±0.260	0.531 ±0.184	13.566 ±5.831	0.423 ±0.176	0.321 ±0.205
EvoGen	0.614 ±0.245	7.122 ±6.065	0.618 ±0.294	0.594 ±0.321	0.496 ±0.195	12.571 ±5.506	0.412 ±0.213	0.336 ±0.245	0.512 ±0.192	14.143 ±5.778	0.390 ±0.175	0.292 ±0.209
EvoDiff	0.680 ±0.229	4.868 ±5.091	0.719 ±0.260	0.691 ±0.293	0.532 ±0.200	11.141 ±5.713	0.478 ±0.220	0.397 ±0.255	0.533 ±0.192	13.116 ±5.835	0.434 ±0.181	0.332 ±0.212
MSAGPT	0.647 ±0.243	5.951 ±5.660	0.670 ±0.287	0.644 ±0.316	0.521 ±0.202	12.121 ±5.940	0.446 ±0.229	0.375 ±0.262	0.597 ±0.204	11.148 ±6.643	0.538 ±0.239	0.439 ±0.276
LatentMSA(3B)	0.678 ±0.221	4.695 ±4.966	0.739 ±0.250	0.714 ±0.275	<u>0.715</u> ±0.198	5.520 ±5.613	0.753 ±0.241	0.689 ±0.266	0.822 ±0.122	3.154 ±3.499	0.886 ±0.135	0.824 ±0.174
LatentMSA(E2E)	**0.733** ±0.208	**3.859** ±4.464	**0.797** ±0.231	**0.776** ±0.254	**0.746** ±0.181	**4.783** ±4.920	**0.789** ±0.229	**0.735** ±0.253	**0.849** ±0.102	**2.544** ±2.883	**0.913** ±0.112	<u>0.867</u> ±0.137

On shared FB_PROTEIN targets, LatentMSA generally agrees better with the UniRef100-derived reference than discrete generation baselines (Table 1). LatentMSA$_{\text{ESM2-3B}}$ achieves the highest all-pair and sequence-separated Spearman correlations (0.646 and 0.637), while LatentMSA$_{\text{ESM2-15B}}$ achieves the highest reference overlap at $L/2$, L, and $2L$ (0.513, 0.533, and 0.543). Scaling is non-monotonic across the evaluated backbones, and Figure 2 shows that LatentMSA does not lead on every target. Overall, the results support higher average coupling fidelity for the continuous latent representations without implying universal target-level dominance. Appendix A.11 examines the contributions of statistical alignment, spatial similarity alignment, and probe diversity.

4.4 Offline LatentMSA is robust across MSA-depth regimes

LatentMSA is trained offline to synthesize fixed-depth pseudo-MSAs from single-sequence ESM representations without target-specific homology retrieval at inference time; concat MSA depth, $d_{\mathrm{UniRef100}} + d_{\mathrm{ColabFoldDB}}$, is used only for post-hoc stratification. As shown in Table 2 and Appendix Table 6, LatentMSA(3B), the offline reference, outperforms EvoGen, EvoDiff, and MSAGPT in nearly all bucket-wise metric comparisons. The only exceptions are LDDT on FB_PROTEIN $[1, 100)$, where EvoDiff is marginally higher (0.680 vs. 0.678), and RMSD on CASP15 $[1, 100)$, where EvoGen is better (13.701 vs. 15.613). On FB_PROTEIN, LatentMSA(3B) exceeds the three discrete MSA-generation baselines on every metric in the $[100, 500)$ and $[500, 16\mathrm{k})$ strata, demonstrating that its advantage extends beyond targets with few retrieved homologs. Nevertheless, the offline model remains behind ESMFold and SimpleFold on most high-depth metrics, indicating that matching MSA statistics alone does not fully optimize downstream structural accuracy.

4.5 Effective Coupling Fidelity is associated with downstream performance

Figure 3 compares three discrete MSA-generation baselines and six offline LatentMSA variants. EvoDiff, EvoGen, and MSAGPT have ECF values of 0.001, 0.053, and 0.130 and generally occupy the low-ECF, lower-accuracy region, although their ordering is non-monotonic. Across all nine methods, Spearman correlations between ECF and TM-score are 0.817 ($p = 0.0072$), 0.883 ($p = 0.0016$), and 0.967 ($p = 2.2 \times 10^{-5}$) for the three increasing-depth buckets; for LDDT, they are 0.467 ($p = 0.205$), 0.883 ($p = 0.0016$), and 0.967 ($p = 2.2 \times 10^{-5}$). The association is strongest at high depth, whereas shallow-depth LDDT shows no significant monotonic association. These model-level correlations and unadjusted two-sided p-values are descriptive; ECF is neither depth-independent nor causal and should not be interpreted as evidence of a mechanistic relationship.

Scaling is non-monotonic under the current recipe. Relative to 3B, 15B slightly improves reference overlap at $K = L$ from 0.523 to 0.533, but lowers sequence-separated Spearman from 0.637 to 0.569 and ECF from 0.499 to 0.434. Its high-depth TM-score and LDDT also decrease from 0.886 and 0.822 to 0.864 and 0.804. All six ESM-2 variants share training hyperparameters and latent-generator configuration, including $d_q = 512$. The fixed interface capacity and optimization settings may therefore be suboptimal for 15B, but the cause remains untested. ECF is thus a compact empirical summary of pseudo-MSA coupling fidelity, not a complete quality measure.

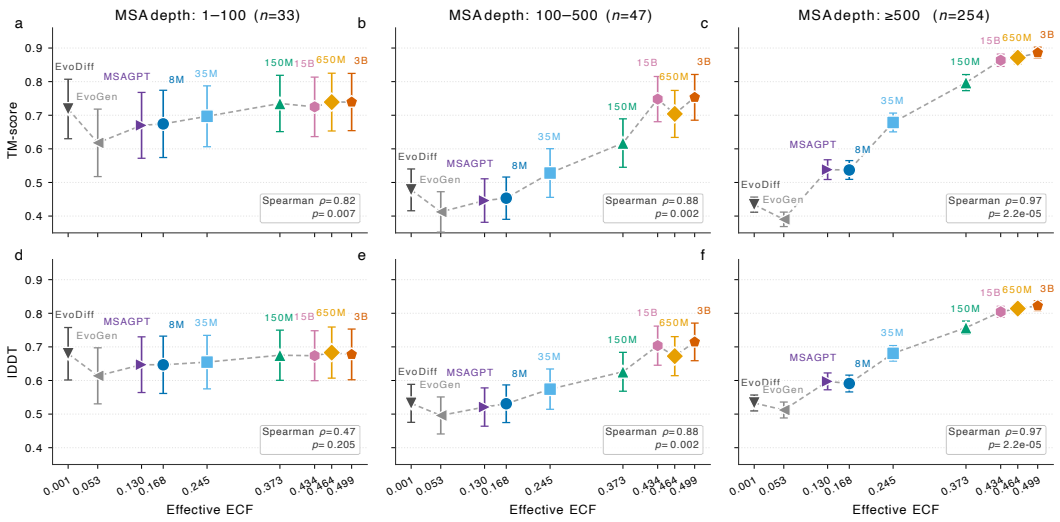


Figure 3: ECF versus downstream structure accuracy across three discrete MSA-generation baselines and six offline LatentMSA variants on FB_PROTEIN. Rows report TM-score and LDDT; columns correspond to concat-MSA-depth buckets $[1, 100)$, $[100, 500)$, and $[500, 16\text{k})$. Points and error bars show bucket means and approximate 95% confidence intervals (mean $\pm$ 1.96 SEM); each panel reports the Spearman correlation and two-sided p-value across the nine methods.

4.6 Structure-supervised adaptation improves most depth-stratified results and closes high-depth gaps

We adapt LatentMSA with the folding objective while keeping the PLM and Protenix frozen. Structural gradients pass through the frozen folding pipeline to update the dynamic anchor, probe directions, gating modules, state-specific self-attention layers, and auxiliary heads. Starting from the offline LatentMSA(3B) weights, we fine-tune this configuration for 2,000 steps on eight GPUs and denote it as LatentMSA(E2E). It improves all twelve bucket-level means over LatentMSA(3B) on FB_PROTEIN. The gains are largest in the $[1, 100)$ stratum for all four metrics, where LatentMSA(E2E) ranks first among all methods. As detailed in Appendix Table 6, this adaptation improves eleven of the twelve bucket-level means on CASP15; the only regression is RMSD in $[100, 500)$ (7.739 vs. 7.546). In the CASP15 $[1, 100)$ stratum, LatentMSA(E2E) achieves the best LDDT, TM-score, and GDT_TS, while EvoGen remains best in RMSD. Structure-supervised adaptation also closes or reverses several high-depth gaps. On FB_PROTEIN $[500, 16\text{k})$, for example, the offline TM-score of 0.886 trails ESMFold's 0.911, whereas LatentMSA(E2E) reaches 0.913; it also achieves better LDDT and RMSD and matches ESMFold's displayed GDT_TS of 0.867. On CASP15 in the same depth stratum, LatentMSA(E2E) obtains the best RMSD and matches ESMFold's displayed TM-score and GDT_TS, although its LDDT remains lower. These results show that structure-supervised LatentMSA adaptation yields broad improvements across concat-MSA-depth regimes, with the largest shallow-depth gains on FB_PROTEIN and more metric-dependent behavior on CASP15.

5 Conclusion

We present LatentFold, which supplies structure-relevant MSA signals as continuous latent states rather than discrete homologous sequences. LatentMSA uses permutation-invariant statistical alignment without structural labels or target-specific retrieval. Its differentiable interface enables structural gradients to update LatentMSA through a frozen folding backbone. Across FoldBench and CASP15, offline LatentFold outperforms discrete MSA-generation baselines in most depth-stratified comparisons; structure-supervised adaptation improves nearly all bucket-level results. ECF is associated with downstream accuracy, but scaling is non-monotonic and the association is neither depth-independent nor causal. Overall, continuous latent MSA generation is a trainable alternative to discrete homolog generation for supplying structure-relevant MSA signals.

AI USE STATEMENT

In this work, we used generative AI tools to assist in refining the presentation of the paper's conceptual framing and the interpretation of existing experimental results. We additionally used these tools to edit portions of the manuscript for clarity and readability; suggest alternative wording, section-level narrative structure, terminology, the paper title, and figure captions; and assist with LaTeX consistency and compilation checks. Generative AI was not used to generate data, conduct experiments, derive the reported numerical results, formulate or prove mathematical claims, or select or fabricate references. All AI-assisted text and suggestions were reviewed and revised by the authors, and all scientific claims were checked against the reported methods, experimental results, and cited literature. The authors take full responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide an anonymous repository containing the core implementation for peer review at `https://anonymous.4open.science/r/LatentFold-2C68`. All datasets used in this work are publicly available; their sources, release cutoffs, filtering criteria, MSA construction, and evaluation protocol are described in Section 4.1 and the appendix. The appendix also provides the model formulation, tensor dimensions, loss definitions, and implementation and training details needed to reproduce the reported experiments. We will release the complete codebase upon publication.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

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. *BioRxiv*, pp. 2023–09, 2023.

Amy C Anderson. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003.

Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096): 223–230, 1973. doi: 10.1126/science.181.4096.223. URL `https://www.science.org/doi/abs/10.1126/science.181.4096.223`.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754. URL `https://www.science.org/doi/abs/10.1126/science.abj8754`.

Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-International Conference on Learning Representations*, 2022.

Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Henry Chao, Li Chen, Paul A Craig, Gregg V Crichlow, Kenneth Dalenberg, Jose M Duarte, Shuchismita Dutta, Maryam Fayazi, Zukang Feng, Justin W Flatt, Sai Ganesan, Sutapa Ghosh, David S Goodsell, Rachel Kramer Green, Vladimir Guranovic, Jeremy Henry, Brian P Hudson, Igor Khokhriakov, Catherine L Lawson, Yuhe Liang, Robert Lowe, Ezra Peisach, Irina Persikova, Dennis W Piehl, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Brinda Vallat, Maria Voigt, Ben Webb, John D

Westbrook, Shamara Whetstone, Jasmine Y Young, Arthur Zalevsky, and Christine Zardecki. RCSB protein data bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.*, 51(D1):D488–D508, January 2023.

Hanqun Cao, Xinyi Zhou, Zijun Gao, Chenyu Wang, Xin Gao, Zhi Zhang, Cesar de la Fuente-Nunez, Chunbin Gu, Ge Liu, and Pheng-Ann Heng. Lightweight msa design advances protein folding from evolutionary embeddings. *arXiv preprint arXiv:2507.07032*, 2025.

Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 37504–37534. Curran Associates, Inc., 2024. doi: 10.52202/079017-1184. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/41f3347f8f47c17bbadaed584e68d8bd-Paper-Conference.pdf.

C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826, 1986. ISSN 1460-2075. doi: 10.1002/j.1460-2075.1986.tb04288.x. URL https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.

Ken A. Dill and Hue Sun Chan. From levinthal to pathways to funnels. *Nature Structural Biology*, 4(1):10–19, 1997. ISSN 1545-9985. doi: 10.1038/nsb0197-10. URL https://doi.org/10.1038/nsb0197-10.

Stanley D Dunn, Lindi M Wahl, and Gregory B Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2008.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pp. 258–267, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

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

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in neural information processing systems*, 25, 2012.

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

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

Felix Kallenborn, Alejandro Chacon, Christian Hundt, Hassan Sirelkhatim, Kieran Didi, Sooyoung Cha, Christian Dallago, Milot Mirdita, Bertil Schmidt, and Martin Steinegger. Gpu-accelerated homology search with mmseqs2. *Nature Methods*, 22(10):2024–2027, 2025.

Taehyeon Kim, Jaehoon Oh, Nak Yil Kim, Sangwook Cho, and Se Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pp. 2628–2635. International Joint Conferences on Artificial Intelligence, 2021.

P E Leopold, M Montal, and J N Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, 89(18):8721–8725, 1992. doi: 10.1073/pnas.89.18.8721. URL https://www.pnas.org/doi/abs/10.1073/pnas.89.18.8721.

Levinthal, Cyrus. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44–45, 1968. doi: 10.1051/jcp/1968650044. URL https://doi.org/10.1051/jcp/1968650044.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International conference on machine learning*, pp. 1718–1727. PMLR, 2015.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PLOS ONE*, 6(12):1–20, 12 2011. doi: 10.1371/journal.pone.0028766. URL https://doi.org/10.1371/journal.pone.0028766.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://doi.org/10.1038/s41592-022-01488-1.

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José 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.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free, 2025. URL https://arxiv.org/abs/2505.06708.

Jason Ramapuram, Federico Danieli, Eeshan Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russ Webb. Theory, analysis, and best practices for sigmoid self-attention. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Zhdhg6n2OG.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International conference on machine learning*, pp. 8844–8856. PMLR, 2021.

Jean-Paul Renaud, Ashwin Chari, Claudio Ciferri, Wen-ti Liu, Hervé-William Rémigy, Holger Stark, and Christian Wiesmann. Cryo-em in drug discovery: achievements, limitations and prospects. *Nature reviews Drug discovery*, 17(7):471–492, 2018.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

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

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

Johannes Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21 (7):951–960, 11 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti125. URL https://doi.org/10.1093/bioinformatics/bti125.

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. doi: 10.1101/2025.01.08.631967. URL https://www.biorxiv.org/content/early/2025/01/11/2025.01.08.631967.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic acids research*, 51(D1):D523–D531, 2023.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

Yuyang Wang, Jiarui Lu, Navdeep Jaitly, Joshua Susskind, and Miguel Angel Bautista. Simplefold: Folding proteins is simpler than you think. In *International Conference on Learning Representations*, volume 2026, pp. 32027–32054, 2026.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022. doi: 10.1101/2022.07.21.500999. URL https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999.

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(1):442, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-67127-3. URL https://doi.org/10.1038/s41467-025-67127-3.

Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The i-tasser suite: protein structure and function prediction. *Nature Methods*, 12(1):7–8, 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3213. URL https://doi.org/10.1038/nmeth.3213.

Jun Zhang, Sirui Liu, Mengyun Chen, Haotian Chu, Min Wang, Zidong Wang, Jialiang Yu, Ningxi Ni, Fan Yu, Dechin Chen, Yi Isaac Yang, Boxin Xue, Lijiang Yang, Yuan Liu, and Yi Qin Gao. Unsupervisedly prompting alphafold2 for accurate few-shot protein structure prediction. *Journal of Chemical Theory and Computation*, 19(22):8460–8471, 2023. doi: 10.1021/acs.jctc.3c00528. URL https://doi.org/10.1021/acs.jctc.3c00528. PMID: 37947474.

Le Zhang, Jiayang Chen, Tao Shen, Yu Li, and Siqi Sun. Msa generation with seqs2seqs pretraining: Advancing protein structure predictions. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 57324–57348. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1827. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/694be3548697e9cc8999d45e8d16fe1e-Paper-Conference.pdf.

Yang Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, 2008. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2008.02.004. URL https://www.sciencedirect.com/science/article/pii/S0959440X08000341. Nucleic acids / Sequences and topology.

Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024b.

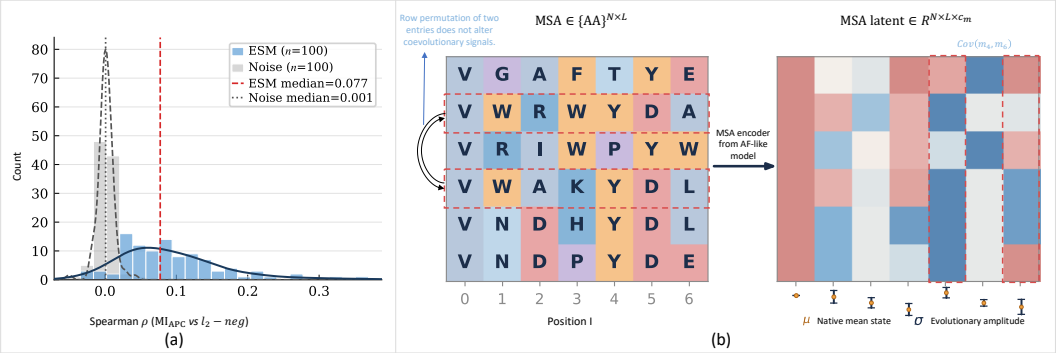


Figure 4: (a) ESM-2 feature similarities show a weak association with MSA-derived pairwise patterns without target-specific retrieval at inference time. (b) Permutation-invariant aggregation avoids explicit row-wise matching: channel moments summarize within-row features, while channel correlations and depth-averaged residue-pair similarities constrain selected dependencies.

A Supplementary Experimental Details

A.1 Empirical Motivation: ESM-2 Latents Contain Weak MSA-Associated Signals

We examined whether single-sequence ESM-2 representations contain signals associated with MSA-derived residue-pair patterns without target-specific homology retrieval at inference time. Prior studies have shown that PLM representations and attention maps contain structure-relevant signals (Rao et al., 2021; Zhang et al., 2024b). We sampled 100 monomer sequences from FoldBench (Xu et al., 2025) and estimated MSA-derived pairwise patterns using mutual information (MI) with average product correction (APC) (Dunn et al., 2008). In parallel, we extracted ESM-2 residue embeddings and constructed a single-sequence pairwise matrix using negative residue-level L_2 distance, as detailed in Appendix A.3.

ESM-2 feature similarities exhibit a weak but consistent correlation with the MSA-derived pairwise patterns. As shown in Figure 4(a), the distribution is right-shifted relative to a dimension-matched random baseline, with median $\rho = 0.077$ compared with $\rho = 0.001$ for random features. This observation indicates a statistical association, but does not by itself establish that ESM-2 recovers family-specific coevolutionary couplings.

This association motivates using single-sequence features to construct a continuous carrier of selected structure-relevant MSA-derived signals rather than recovering individual homologous sequences. LatentMSA uses sequence-conditioned probes and state-specific self-attention to produce distinct latent rows, which are subsequently constrained by selected natural-MSA-derived statistics. These rows are probe-conditioned latent states, not recovered homologs or samples from the complete natural-MSA distribution.

A.2 Coupling Matrix for Discrete MSA Tokens: Mutual Information with APC

For analyses performed directly in discrete token space, including the empirical-motivation analysis in Appendix A.1, we compute pairwise mutual information (MI) with pseudocount regularization and average product correction (APC) as a coupling score. This token-space construction is not used for the cross-method coupling comparison in Table 1; that comparison uses the common projected-space procedure described in Appendix A.3.

Algorithm Details:

1. **One-hot Encoding:** The tokens are first one-hot encoded. The number of classes is defined as $q = \min(\text{max_token} + 1, 22)$ (representing the 20 standard amino acids, a gap token, and an unknown token). The MSA is represented as a tensor $\mathbf{X} \in \{0,1\}^{N\times L\times q}$, where $X[n,l,a] = 1$ if and only if row n contains amino acid a at column l.

2. **Single-site Frequencies:** We compute the marginal probability of observing amino acid a at position i with a pseudocount $pc = \frac{\text{pseudocount}}{q}$ (defaulting to $\frac{0.5}{q}$):

$$f_i(a) = \left(\frac{1}{N} \sum_{n=1}^{N} X[n, i, a] \right) + pc.$$

The frequencies are then renormalized such that $\sum_a f_i(a) = 1$.

3. **Pairwise Frequencies:** For every pair of positions $i < j$, the joint probability of observing amino acids a and b is computed as:

$$f_{ij}(a, b) = \left(\frac{1}{N} \sum_{n=1}^{N} X[n, i, a] \cdot X[n, j, b] \right) + \frac{pc}{q}.$$

This is also renormalized to ensure $\sum_{a,b} f_{ij}(a, b) = 1$.

4. **MI:** The raw mutual information between positions i and j is calculated as:

$$\mathrm{MI}(i, j) = \sum_{a,b} f_{ij}(a, b) \log \frac{f_{ij}(a, b)}{f_i(a) f_j(b)}.$$

The summation is strictly applied to entries where both the joint probability and the outer product of marginal probabilities are positive.

5. **APC:** To suppress background noise arising from entropy and phylogenetic biases, we apply the APC:

$$\mathrm{MI}^{\mathrm{APC}}(i, j) = \mathrm{MI}(i, j) - \frac{\overline{\mathrm{MI}}(i, \cdot)\overline{\mathrm{MI}}(\cdot, j)}{\overline{\mathrm{MI}}(\cdot, \cdot)},$$

where $\overline{\mathrm{MI}}(i, \cdot)$ denotes the row mean over j, and $\overline{\mathrm{MI}}(\cdot, \cdot)$ is the global mean. The diagonal elements are zeroed out. The final resulting matrix is denoted as $\mathbf{M}_{\mathrm{MI}} \in \mathbb{R}^{L \times L}$.

A.3 Coupling Matrix for Continuous Embeddings: Negative L_2 Distance

For the cross-method comparison in Table 1, the natural MSA and each discrete MSA produced by EvoDiff, EvoGen, or MSAGPT are processed by the same frozen MSA linear projection from the structure prediction model, yielding continuous tensors in a shared feature space. Because this projection was learned within the structure predictor, its output space is adapted to structure-relevant MSA features rather than being an arbitrary embedding space. LatentMSA is trained to generate pseudo-MSA tensors directly in the same space. Applying the identical projection and coupling computation to all MSA sources makes their rank agreement and reference-coupling overlap directly comparable within this representation space.

For these continuous representations (e.g., $\mathbf{M}_{\text{true}}$ and $\mathbf{M}_{\text{pseudo}}$), computing MI over binned real-valued vectors is statistically noisy and computationally expensive. We instead use negative pairwise Euclidean distance as a representation-derived coupling surrogate for comparing rank structure; larger values indicate greater similarity under this surrogate and should not be interpreted directly as physical coupling strength.

Depending on the input shape regime, the computation is defined as follows:

2D Input (e.g., ESM per-residue embeddings $\mathbf{h} \in \mathbb{R}^{L \times D}$):

When processing a single query sequence, the representation simplifies directly to:

$$S(i, j) = -\|\mathbf{h}_i - \mathbf{h}_j\|_2.$$

3D Input (e.g., MSA latent features $\mathbf{M} \in \mathbb{R}^{N \times L \times C}$):

For every MSA sequence n, we compute the full pairwise L_2 distance matrix using the norm expansion $\|a-b\|^2 = \|a\|^2 + \|b\|^2 - 2a \cdot b$. The negative distance is then averaged across all N sequences:

$$S(i,j) = -\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{M}[n,i,:] - \mathbf{M}[n,j,:]\|_2.$$

We average across MSA rows after computing each row's distance matrix rather than mean-pooling the embeddings before distance calculation. This preserves within-row variation that would otherwise be removed by pre-pooling. In both the 2D and 3D settings, diagonal entries are set to zero.

A.4 Coupling-Fidelity Metrics and Effective Coupling Fidelity

Let $\mathbf{S}^{\text{pred}}, \mathbf{S}^{\text{ref}} \in \mathbb{R}^{L\times L}$ denote the coupling matrices of a generated representation and its natural-MSA reference. We evaluate only non-redundant upper-triangular residue pairs. Define

$$\Omega_{\text{all}} = \{(i,j) : 1 \le i < j \le L\}, \qquad \Omega_{\text{LR}} = \{(i,j) \in \Omega_{\text{all}} : |i-j| \ge 6\}.$$

Rank agreement. Spearman (all) is the Spearman rank correlation between the entries of $\mathbf{S}^{\text{pred}}$ and $\mathbf{S}^{\text{ref}}$ indexed by Ω_{all}. Spearman (LR), denoted ρ_{LR}, is computed analogously over Ω_{LR}. Thus, these metrics compare coupling order rather than requiring the two matrices to share an absolute numerical scale; the LR variant excludes local sequence neighbors.

Reference-coupling overlap. For $K \in \{\lfloor L/2 \rfloor, L, 2L\}$, let $K' = \min(K, |\Omega_{\text{LR}}|)$ and let $\mathcal{T}_{K'}(\mathbf{S})$ contain the K' highest-scoring pairs in Ω_{LR}. We compute

$$p_{\text{Top-K}} = \frac{|\mathcal{T}_{K'}(\mathbf{S}^{\text{pred}}) \cap \mathcal{T}_{K'}(\mathbf{S}^{\text{ref}})|}{K'}.$$

Reference overlap at $L/2$, L, and $2L$ uses $K = \lfloor L/2 \rfloor$, L, and $2L$, respectively. These metrics quantify overlap between ranked coupling-score sets and should not be interpreted as contact precision or contact recovery against an experimentally determined structure.

Effective Coupling Fidelity. For model-level reporting, let $\bar{\rho}_{\text{LR}}$ and $\bar{p}_{\text{Top-L}}$ be the corresponding means over the shared evaluation targets. We first define their joint fidelity

$$q_{\text{eff}} = \bar{\rho}_{\text{LR}}\, \bar{p}_{\text{Top-L}}, \qquad \text{ECF} = \frac{q_{\text{eff}}}{1 - q_{\text{eff}}}. \tag{10}$$

The product acts as a soft conjunction: ρ_{LR} measures ranking consistency over sequence-separated pairs, whereas $p_{\text{Top-L}}$ measures reference-coupling overlap at $K = L$. Consequently, q_{eff} is high only when both criteria are high. The monotonic odds-like transformation increases separation at high joint fidelity but introduces no independent ranking information. ECF is therefore an empirical coupling-fidelity index rather than a calibrated probability or physical measurement.

For target-level analyses, the same transformation is applied to target-specific ρ_{LR} and reference-coupling overlap at $K = L$ only when both lie in $[0, 1]$; negative correlations are treated as undefined rather than clipped or converted to absolute values. In the shared evaluation set, only EvoDiff has negative ρ_{LR}, on three targets; all target-level values for the other methods are non-negative. These three correlations remain included in EvoDiff's model-level mean $\bar{\rho}_{\text{LR}}$ but are excluded from target-level ECF ranking. Finally, the model-level score uses $\mathbb{E}[\rho_{\text{LR}}]\mathbb{E}[p_{\text{Top-L}}]$, which need not equal $\mathbb{E}[\rho_{\text{LR}} p_{\text{Top-L}}]$. The two inputs may also be statistically dependent and are given equal importance by construction. ECF should therefore be interpreted as a compact model-level summary of coupling fidelity rather than a complete measure of pseudo-MSA quality or a causal predictor of folding accuracy.

A.5 Qualitative Folding-Landscape Analogy for the Latent Anchor

Folding landscapes provide a qualitative analogy for representing multiple states around a sequence-conditioned reference (Anfinsen, 1973; Levinthal, Cyrus, 1968; Leopold et al., 1992; Dill & Chan,

1997). In LatentFold, the anchor $\mathbf{e}_{\text{anchor}}$ is projected from a global PLM-derived sequence representation and serves only as the center of the generated probe set. Perturbations around this anchor can be viewed heuristically as exploring nearby latent states. LatentFold does not estimate free energies, thermodynamic probabilities, physical transition paths, or a native-state energy minimum; the analogy is not a mechanistic interpretation of the learned representation.

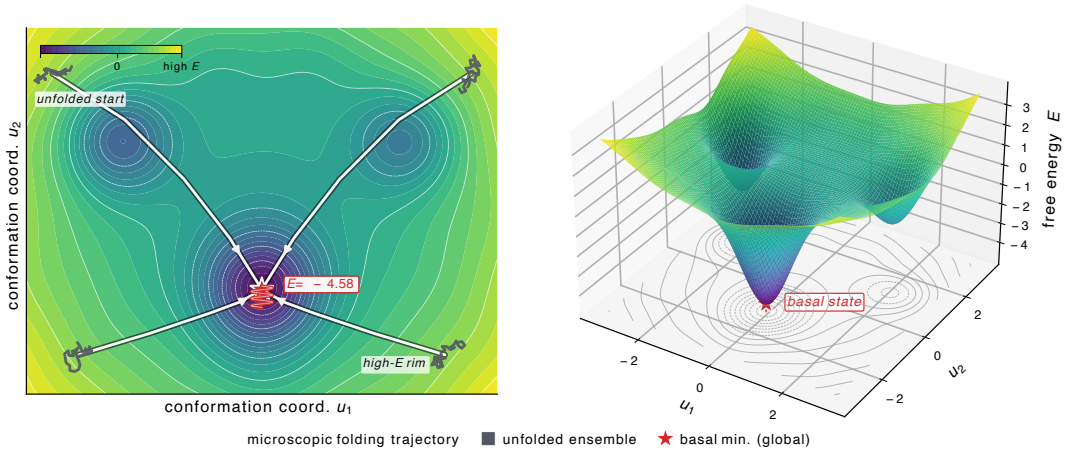


Figure 5: A folding-landscape schematic used only as a qualitative analogy for generating multiple latent probes around a sequence-conditioned reference. The landscape is not estimated by LatentFold and does not assign thermodynamic meaning to the learned anchor.

A.6 Full Tensor Dimensions and Computational Graph of LatentFold

Figure 6 shows the detailed computational graph and tensor flow of the LatentFold framework, including the forward pass and corresponding optimization objectives.

(1) Latent Generator & Decoupled Perturbation: A single input sequence is processed by a frozen ESM-2 encoder to extract per-residue embeddings $\mathbf{H}_{\text{esm}} \in \mathbb{R}^{L \times d_{\text{esm}}}$. Global pooling followed by an MLP projects these features into a sequence-specific dynamic anchor $\mathbf{e}_{\text{anchor}} \in \mathbb{R}^{d_q}$. To generate a set of sequence-conditioned latent states, N probes $\mathbf{E}_{\text{probes}} \in \mathbb{R}^{N \times d_q}$ are formed through decoupled perturbations, which separate normalized direction vectors $\mathbf{P}/\|\mathbf{P}\|$ from the learnable scale λ_{radius}.

(2) Modulation & State-Specific Attention: The latent probes $\mathbf{E}_{\text{probes}}$ interact with the sequence representations through sigmoid gating, producing a soft modulation mask $\mathbf{A}_{\text{mask}}$. This mask modulates the projected value features and conditions the subsequent State-Specific Self-Attention module. The attention module then produces the pseudo-MSA representation $\mathbf{M}_{\text{pseudo}} \in \mathbb{R}^{N \times L \times c_m}$. Additional residue-level heads aggregate this representation to predict the MSA profile and deletion mean, supervised by $\mathcal{L}_{\text{profile}}$ and $\mathcal{L}_{\text{del}}$, respectively.

(3) Decoupled Manifold Alignment Loss: To supervise latent representation construction without discrete row-wise sequence matching, a reference representation $\mathbf{M}_{\text{true}} \in \mathbb{R}^{N_t \times L \times c_m}$ is obtained from natural MSAs through a frozen MSA projection module. LatentFold aligns selected depth-aggregated statistics of $\mathbf{M}_{\text{pseudo}}$ and $\mathbf{M}_{\text{true}}$, including the mean ($\mathcal{L}_{\text{mean}}$), standard deviation ($\mathcal{L}_{\text{std}}$), and channel/residue-pair similarity structures ($\mathcal{L}_{\text{corr}}$).

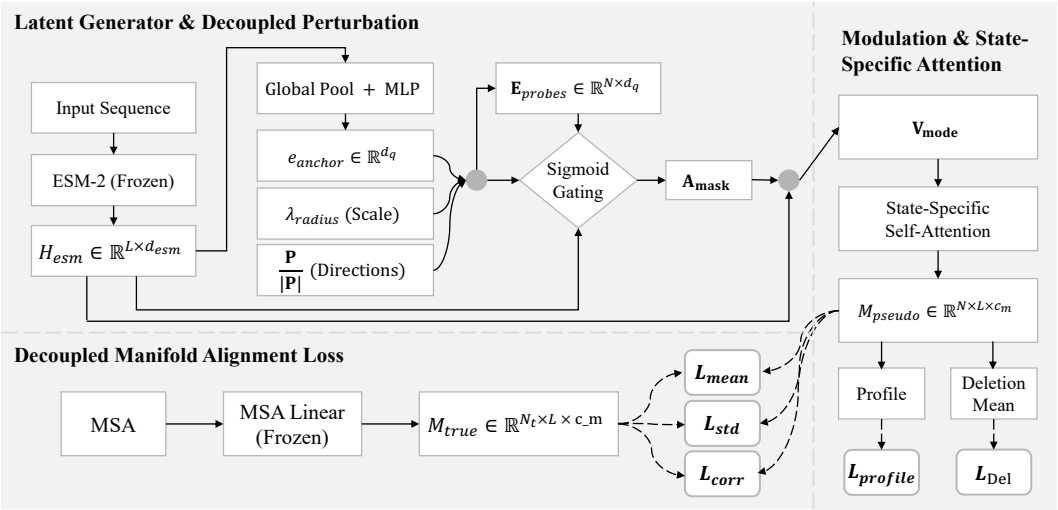


Figure 6: The LatentFold architecture. Rather than generating discrete homologous sequences, a frozen PLM and LatentMSA construct an MSA-shaped collection of continuous, probe-conditioned latent states ($\mathbf{M}_{\text{pseudo}}$). The Decoupled Manifold Alignment Loss trains this representation by matching selected depth-aggregated moments and similarity statistics with a natural-MSA-derived reference ($\mathbf{M}_{\text{true}}$), without explicit row-wise correspondence.

A.7 Detailed Formulation of Decoupled Manifold Alignment Loss

This section defines the statistical terms used in the Decoupled Manifold Alignment Loss. For each latent row $\mathbf{m}_n \in \mathbb{R}^{L \times c_m}$, we compute channel-wise empirical statistics over residue positions:

$$\begin{aligned} \mu_n &= \frac{1}{L} \sum_{l=1}^{L} \mathbf{m}_{n,l} \in \mathbb{R}^{c_m}, \\ \sigma_n &= \sqrt{\frac{1}{L} \sum_{l=1}^{L} (\mathbf{m}_{n,l} - \mu_n)^2 + \epsilon} \in \mathbb{R}^{c_m}. \end{aligned} \tag{11}$$

The normalized representation is

$$\mathbf{Z}_n = (\mathbf{m}_n - \mu_n) \oslash \sigma_n, \tag{12}$$

where $\oslash$ denotes element-wise division and ϵ is a small constant for numerical stability.

Mean and standard-deviation alignment. The mean loss aligns the central tendency of the generated and natural-MSA latent representations:

$$\mathcal{L}_{\text{mean}} = \left\| \mathbb{E}_n[\mu_n^{\text{pseudo}}] - \mathbb{E}_n[\mu_n^{\text{true}}] \right\|_1. \tag{13}$$

The standard deviation loss is computed in log-space to stabilize optimization and improve sensitivity to small-variance channels:

$$\mathcal{L}_{\text{std}} = \left\| \log\left(\mathbb{E}_n[\sigma_n^{\text{pseudo}}] + \epsilon\right) - \log\left(\mathbb{E}_n[\sigma_n^{\text{true}}] + \epsilon\right) \right\|_1. \tag{14}$$

Channel correlation alignment. To match channel-wise dependency structures, we compute the channel correlation matrix for each latent row:

$$\mathbf{C}_n = \frac{1}{L} \mathbf{Z}_n^\top \mathbf{Z}_n \in \mathbb{R}^{c_m \times c_m}. \tag{15}$$

The channel correlation loss is

$$\mathcal{L}_{\text{corr}}^{C} = \left\| \mathbb{E}_n[\mathbf{C}_n^{\text{pseudo}}] - \mathbb{E}_n[\mathbf{C}_n^{\text{true}}] \right\|_1. \tag{16}$$

Spatial similarity alignment. To compare residue-level similarity structure, we compute a row-level matrix and then average it across depth:

$$\mathbf{R} = \mathbb{E}_n[\mathbf{R}_n], \qquad \mathbf{R}_n = \mathrm{Norm}(\mathbf{Z}_n)\,\mathrm{Norm}(\mathbf{Z}_n)^\top \in \mathbb{R}^{L\times L}, \tag{17}$$

where $\mathrm{Norm}(\cdot)$ denotes row-wise L_2 feature normalization. The spatial similarity loss is

$$\mathcal{L}^L_{\text{corr}} = \left\|\mathbf{R}^{\text{pseudo}} - \mathbf{R}^{\text{true}}\right\|_1. \tag{18}$$

In practice, this term is masked to sequence-separated residue pairs above a predefined threshold, reducing the contribution of local sequential similarity. The threshold and mask are applied identically to generated and natural-MSA latent representations.

The final correlation loss combines channel and spatial terms:

$$\mathcal{L}_{\text{corr}} = \mathcal{L}^C_{\text{corr}} + \alpha_{\text{spatial}}\mathcal{L}^L_{\text{corr}}. \tag{19}$$

Probe diversity regularization. Moment alignment constrains global distributional statistics, but it does not by itself prevent all generated probes from collapsing to redundant latent states. We therefore regularize latent probes using the margin-based repulsion loss:

$$\mathcal{L}_{\text{div}} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\max\left(0, \cos(\mathbf{E}_i, \mathbf{E}_j) - \tau_{\max}\right). \tag{20}$$

Total objective. The final training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mean}} + \lambda_{\text{std}}\mathcal{L}_{\text{std}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \mathcal{L}_{\text{aux}}, \tag{21}$$

where $\mathcal{L}_{\text{aux}}$ includes auxiliary profile and deletion-mean prediction losses, which are described in Appendix A.8. Since all distributional terms are aggregated symmetrically over the depth dimension, the objective does not require row-wise correspondence between generated and natural MSAs.

A.8 Auxiliary Feature Prediction for AF3 Compatibility

To integrate with AF3-style downstream structure prediction modules, LatentMSA includes two residue-level prediction heads at its output. They predict the MSA token-profile distribution and deletion-mean feature used to construct downstream *s_inputs*.

A.8.1 Residue-Level Aggregation and Prediction Architecture

Given the pseudo-MSA tensor $\mathbf{M}_{\text{pseudo}} \in \mathbb{R}^{N\times L\times c_m}$, we average over the probe-depth dimension to obtain $\mathbf{H}_{res} \in \mathbb{R}^{L\times c_m}$. A shared nonlinear multilayer perceptron with LayerNorm and GELU activations then feeds two task-specific branches:

Profile Prediction Head: A linear projection outputs 32-dimensional logits over the residue/token categories used by the downstream model.

Deletion Mean Prediction Head: A scalar output followed by Softplus guarantees a non-negative predicted deletion-mean feature.

A.8.2 Auxiliary Statistical Feature Losses

To co-optimize the profile and deletion-mean predictions, we use task-specific auxiliary losses with sequence-length masking over valid residues.

Profile KL Divergence Loss ($\mathcal{L}_{profile}$): Because the ground-truth MSA profile is a probability distribution over residue categories, we use Kullback–Leibler (KL) divergence between the predicted log-probabilities and the target profile distribution.

Robust Deletion Mean Regression Loss ($\mathcal{L}_{del}$): MSA deletion-mean values often exhibit a long-tailed distribution with a large dynamic range. To reduce the influence of extreme values, we apply logarithmic compression, $\log(1+x)$, to both predicted and target values before regression. We then use a Smooth L1 loss with $\beta = 0.5$ for stable optimization.

The predicted profile and deletion mean, together with the generated latent MSA, are integrated into the downstream structure prediction module (Appendix Algorithms 1 and 2). The complete path is differentiable. In the current structure-supervised experiments, the PLM and Protenix backbones are frozen, and structural-loss gradients update only LatentMSA. The interface could additionally support joint fine-tuning of the folding backbone in future work.

Algorithm 1 Construct an initial 1D embedding (Adapted for AF3-like models)

Require: Raw atom features $\{\mathbf{f}^*\}$, residue types $\mathbf{f}^{\text{restype}}$
Require: **Predicted** profile $\hat{\mathbf{f}}^{\text{profile}}$, **Predicted** deletion mean $\hat{\mathbf{f}}^{\text{deletion_mean}}$ from LatentFold

function InputFeatureEmbedder($\{\mathbf{f}^*\}, \mathbf{f}^{\text{restype}}, \hat{\mathbf{f}}^{\text{profile}}, \hat{\mathbf{f}}^{\text{deletion_mean}}$)
  # Embed per-atom features (Standard AF3 operation).
  $\{\mathbf{a}_i\}, _, _, _ \leftarrow$ AtomAttentionEncoder($\{\mathbf{f}^*\}, \emptyset, \emptyset, \emptyset, c_{\text{atom}} = 128, c_{\text{atompair}} = 16, c_{\text{token}} = 384$)
  # Concatenate the per-token features.
  # Replace MSA-derived features with differentiably predicted latent features.
  $\mathbf{s}_i \leftarrow \text{concat}(\mathbf{a}_i, \mathbf{f}_i^{\text{restype}}, \hat{\mathbf{f}}_i^{\text{profile}}, \hat{\mathbf{f}}_i^{\text{deletion_mean}})$
  return $\{\mathbf{s}_i\}$
end function

Algorithm 2 MSA Module (Adapted for AF3-like models)

Require: **Predicted** latent MSA $\hat{\mathbf{m}}_{si}$ (from our DynamicLatentMSAGenerator)
Require: Pair representation $\{\mathbf{z}_{ij}\}$, Single representation $\{\mathbf{s}_i^{\text{inputs}}\}$
Require: $N_{\text{block}} = 4$, $c_m = 64$

function MsaModule($\{\hat{\mathbf{m}}_{si}\}, \{\mathbf{z}_{ij}\}, \{\mathbf{s}_i^{\text{inputs}}\}, N_{\text{block}}, c_m$)
  # 1: $\mathbf{m}_{Si} = \text{concat}(\mathbf{f}_{Si}^{\text{msa}}, \mathbf{f}_{Si}^{\text{has_deletion}}, \mathbf{f}_{Si}^{\text{deletion_value}})$
  # 2: $\{s\} = \text{SampleRandomWithoutReplacement}(\{S\})$
  # 3: Bypass raw MSA concatenation, sampling, and projection. Inject Latent MSA directly.
  $\mathbf{m}_{si} \leftarrow \hat{\mathbf{m}}_{si}$ ▷ $\mathbf{m}_{si} \in \mathbb{R}^{N_{pseudo} \times L \times c_m}$
  $\mathbf{m}_{si} \leftarrow \mathbf{m}_{si} + \text{LinearNoBias}(\{\mathbf{s}_i^{\text{inputs}}\})$
  for $l \in [1, \ldots, N_{\text{block}}]$ **do**
    # Communication
    $\{\mathbf{z}_{ij}\} \leftarrow \{\mathbf{z}_{ij}\} + \text{OuterProductMean}(\{\mathbf{m}_{si}\})$
    # MSA stack
    $\mathbf{u}_{si} \leftarrow \text{MSAPairWeightedAveraging}(\{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}, c = 8)$
    $\{\mathbf{m}_{si}\} \leftarrow \{\mathbf{m}_{si}\} + \text{DropoutRowwise}_{0.15}(\mathbf{u}_{si})$
    $\{\mathbf{m}_{si}\} \leftarrow \{\mathbf{m}_{si}\} + \text{Transition}(\{\mathbf{m}_{si}\})$
    # Pair stack
    $\{\mathbf{z}_{ij}\} \leftarrow \{\mathbf{z}_{ij}\} + \text{DropoutRowwise}_{0.25}(\text{TriangleMultiplicationOutgoing}(\{\mathbf{z}_{ij}\}))$
    $\{\mathbf{z}_{ij}\} \leftarrow \{\mathbf{z}_{ij}\} + \text{DropoutRowwise}_{0.25}(\text{TriangleMultiplicationIncoming}(\{\mathbf{z}_{ij}\}))$
    $\{\mathbf{z}_{ij}\} \leftarrow \{\mathbf{z}_{ij}\} + \text{DropoutRowwise}_{0.25}(\text{TriangleAttentionStartingNode}(\{\mathbf{z}_{ij}\}))$
    $\{\mathbf{z}_{ij}\} \leftarrow \{\mathbf{z}_{ij}\} + \text{DropoutColumnwise}_{0.25}(\text{TriangleAttentionEndingNode}(\{\mathbf{z}_{ij}\}))$
    $\{\mathbf{z}_{ij}\} \leftarrow \{\mathbf{z}_{ij}\} + \text{Transition}(\{\mathbf{z}_{ij}\})$
  end for
  return $\{\mathbf{z}_{ij}\}$
end function

A.9 Implementation and Training Details

Offline LatentMSA training was conducted with distributed data parallelism on a single server equipped with eight NVIDIA H20 GPUs. The model was trained using a per-GPU batch size of 1 and gradient accumulation over 16 steps, resulting in an effective batch size of 128 samples across all GPUs. Training was run for 60,000 optimization steps.

The optimizer was configured with an initial learning rate of 5×10^{-4}, a minimum learning rate of 1×10^{-6}, and weight decay of 0.01. We used $3{,}000$ warmup steps at the beginning of training. For the topology-alignment stage, we used a restart warmup of $1{,}000$ steps with a restart factor of 0.2, followed by a correlation-weight ramp over $3{,}000$ steps. Checkpoints were saved every $1{,}000$ steps.

For data construction, we used the preprocessed mmCIF/MSA training directory, enabled cached MSA loading, and followed the data-augmentation strategy described in Section 4.1. The maximum number of natural MSA rows used for supervision was capped at 512.

For the latent generator, the generated pseudo-MSA depth was set to $N = 256$. The query dimension was $d_q = 512$, with 8 attention heads, giving a per-head dimension of 64. We used 2 state-specific self-attention layers. To reduce memory overhead, both the depth-wise computation and self-attention computation were chunked with a chunk size of 64. The full training script and configuration are released with the code repository.

During structure-supervised LatentMSA adaptation, we use the Protenix structural-loss framework with the bond-length and PAE terms disabled. The active objective combines the remaining diffusion components (including atom-MSE with DNA/RNA/ligand up-weighting and smooth-LDDT), distogram cross-entropy, and the enabled confidence-head losses, with top-level weights

$$(\alpha_{\text{diff}}, \alpha_{\text{dist}}, \alpha_{\text{conf}}) = \left(4.0,\ 3 \times 10^{-2},\ 1 \times 10^{-4}\right).$$

Both the ESM-2 encoder and Protenix backbone are frozen: the structural loss is evaluated through the complete differentiable folding pipeline, but optimizer updates are applied only to LatentMSA, including its generator, gating and state-specific attention modules, and auxiliary heads.

A.10 MSA Retrieval and Generation Time

We compare per-target MSA retrieval or generation time on 20 CASP14 monomer targets (Table 3). LatentMSA, MSAGPT, and EvoDiff were measured on an NVIDIA A100 GPU. MMseqs2 CPU and GPU-search times are taken from Kallenborn et al. (2025); because these values come from a separate study and hardware setting, the cross-source comparison is descriptive rather than a controlled hardware benchmark. EvoGen is omitted because its MindSpore-GPU dependency could not be installed correctly in our hardware environment. The results nevertheless show a large practical difference: median times are 0.056 s for LatentMSA, 6.350 s for GPU MMseqs2, 57.050 s for CPU MMseqs2, 79.000 s for MSAGPT, and 109.797 s for EvoDiff.

A.11 Ablation Study

We conducted an ablation study on 100 randomly sampled FB_PROTEIN targets. As shown in Figure 7, the results support contributions from three components under this evaluation. First, replacing the separated standard-deviation and correlation losses with a coupled Gram-matrix MSE is associated with less stable optimization, more redundant probe representations, and reference-coupling overlap at $K = L$ of approximately 0.1. Second, removing the $L \times L$ spatial similarity term reduces the median Spearman correlation over sequence-separated pairs from approximately 0.55 to 0.40, supporting its role in matching the long-range reference-MSA coupling pattern. Finally, without margin-based probe repulsion, probe representations become more similar. In contrast, the full LatentFold model produces less redundant probes and higher agreement with the reference coupling pattern.

A.12 Dataset Homology Distribution and Selection Rationale

Table 4 summarizes concat MSA depth across CAMEO, FB_PROTEIN (FoldBench), and CASP15, where $d_{\text{concat}} = d_{\text{UniRef100}} + d_{\text{ColabFoldDB}}$. CAMEO is the most concentrated in the high-depth bucket, with 159 of 194 targets (82.0%) at depth 500 or above, compared with 254 of 334 (76.0%) for FB_PROTEIN and 47 of 65 (72.3%) for CASP15. Below depth 100, the corresponding counts are 16 (8.2%), 33 (9.9%), and 10 (15.4%), respectively.

Rationale for prioritizing FB_PROTEIN over CAMEO. Although CAMEO is a standard community benchmark, we prioritize FB_PROTEIN because it is larger and contains more shallow targets in absolute count. In the lowest concat-depth bucket, FB_PROTEIN contains 11 targets compared

Table 3: Per-target MSA retrieval or generation time (seconds) on CASP14 monomers. Lower is better. LatentMSA, MSAGPT, and EvoDiff were measured on an A100; MMseqs2 values are from Kallenborn et al. (2025).

Target	LatentMSA	MMseqs2	GPU MMseqs2	MSAGPT	EvoDiff
T1027	0.0638	63.44	6.44	78.46	93.7044
T1029	0.0567	55.91	5.11	80.46	69.4898
T1031	0.0539	48.69	5.26	78.44	55.1083
T1033	0.0542	33.87	5.07	82.59	60.3338
T1037	0.1161	57.94	9.99	37.02	252.6227
T1038	0.0566	54.83	7.60	72.81	115.9827
T1039	0.0488	45.51	6.26	73.68	98.1793
T1040	0.0473	37.49	5.41	71.01	100.0059
T1041	0.0532	43.15	7.76	67.33	143.5834
T1042	0.0562	56.86	7.79	52.17	169.7324
T1043	0.0493	47.07	5.51	67.77	108.4809
T1047s1	0.0529	56.07	10.12	66.19	138.9251
T1049	0.0486	57.24	5.68	79.54	103.0657
T1061	0.1102	135.36	68.47	95.98	560.8655
T1064	0.0476	59.36	5.25	163.55	61.1824
T1070	0.1025	65.96	12.50	126.34	200.9844
T1074	0.0689	59.93	6.08	153.64	111.1127
T1090	0.0683	64.21	6.12	146.65	103.3048
T1094	0.1278	91.85	29.61	94.24	296.0477
T1096	0.1260	62.61	8.79	133.38	278.1626
Mean	0.0704	59.87	11.24	91.06	156.0437
Median	0.0564	57.05	6.35	79.00	109.7968

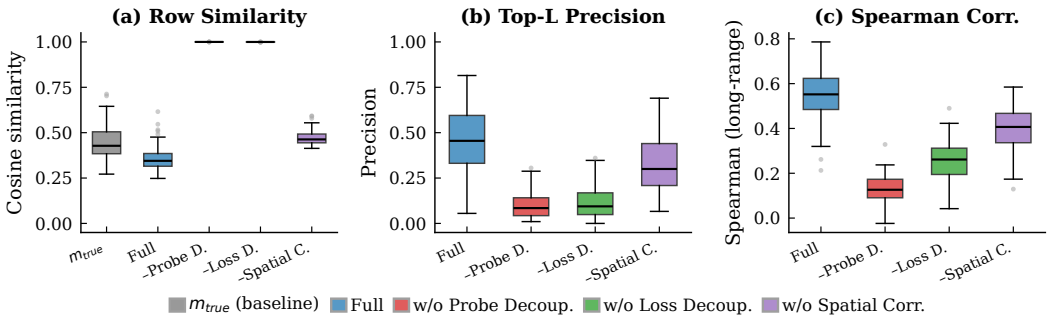


Figure 7: **(a)** Without probe or loss decoupling, latent probes become nearly identical, whereas the full model better matches the diversity of $\mathbf{M}_{\mathrm{true}}$. **(b, c)** These ablations reduce reference-coupling overlap at $K = L$ and Spearman agreement over sequence-separated pairs; removing spatial similarity further reduces agreement with the reference-MSA coupling pattern.

with 5 in CAMEO; across all targets below depth 100, the corresponding counts are 33 and 16. Because high-depth targets still dominate all three datasets, we report depth-stratified results rather than relying only on aggregate averages.

A.13 Detailed Best-of-25 and Ranking-Score Top-1 Results

Table 5 reports ranking-score Top-1 results on FB_PROTEIN, while Tables 6 and 7 report best-of-25 and ranking-score Top-1 results on CASP15, respectively. For ranking-based selection, we retain the highest-scoring prediction among the 25 samples for each target and method.

Table 4: Chain counts and percentages by concat MSA depth, $d_{\mathrm{UniRef100}} + d_{\mathrm{ColabFoldDB}}$. Bold marks each dataset's dominant bucket.

Depth range	CAMEO		FB_PROTEIN		CASP15	
	#chains	%	#chains	%	#chains	%
$[0, 10)$	5	2.6	11	3.3	2	3.1
$[10, 50)$	3	1.5	16	4.8	7	10.8
$[50, 100)$	8	4.1	6	1.8	1	1.5
$[100, 500)$	19	9.8	47	14.1	8	12.3
$[500, \infty)$	**159**	**82.0**	**254**	**76.0**	**47**	**72.3**
Total	194	100.0	334	100.0	65	100.0

Table 5: Per-PDB ranking-score Top-1 comparison grouped by MSA pairing depth on FB_PROTEIN. For each (PDB, method), we select the highest-ranking prediction across all seeds and samples, then average its metrics within each depth bucket. ESMFold and SimpleFold are omitted because their scoring functions differ from the Protenix-based methods, and we do not optimize or cross-calibrate ranking functions. Arrows indicate metric direction; the best method in each (bucket, metric) is bolded.

Method	[1,100), N=33				[100,500), N=47				[500,16k), N=254			
	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑
PX(NO MSA)	0.641 ±0.243	6.462 ±6.358	0.658 ±0.297	0.636 ±0.317	0.503 ±0.205	14.010 ±6.403	0.415 ±0.232	0.348 ±0.255	0.499 ±0.182	16.385 ±6.428	0.365 ±0.173	0.279 ±0.196
EvoGen	0.579 ±0.251	9.435 ±7.553	0.562 ±0.320	0.538 ±0.328	0.463 ±0.195	15.083 ±6.353	0.361 ±0.214	0.294 ±0.234	0.478 ±0.190	16.827 ±6.335	0.334 ±0.172	0.250 ±0.197
EvoDiff	0.647 ±0.238	6.398 ±6.499	0.664 ±0.292	0.642 ±0.313	0.497 ±0.198	14.030 ±6.960	0.426 ±0.231	0.355 ±0.254	0.499 ±0.192	16.113 ±6.549	0.368 ±0.181	0.284 ±0.202
MSAGPT	0.613 ±0.249	7.971 ±7.151	0.606 ±0.313	0.587 ±0.331	0.491 ±0.199	14.447 ±6.598	0.400 ±0.239	0.336 ±0.258	0.564 ±0.207	13.801 ±7.659	0.478 ±0.250	0.394 ±0.277
LatentMSA (8M)	0.604 ±0.264	7.996 ±7.340	0.604 ±0.327	0.583 ±0.343	0.495 ±0.202	14.943 ±8.068	0.402 ±0.239	0.342 ±0.259	0.555 ±0.207	14.367 ±7.924	0.471 ±0.241	0.381 ±0.263
LatentMSA (35M)	0.627 ±0.235	6.868 ±6.298	0.661 ±0.282	0.630 ±0.301	0.536 ±0.213	13.607 ±8.387	0.470 ±0.271	0.405 ±0.291	0.649 ±0.195	10.208 ±7.462	0.625 ±0.250	0.537 ±0.275
LatentMSA (150M)	0.639 ±0.231	6.610 ±6.168	0.676 ±0.278	0.646 ±0.288	0.591 ±0.204	10.862 ±7.648	0.561 ±0.263	0.488 ±0.280	0.729 ±0.167	7.263 ±7.168	0.749 ±0.220	0.671 ±0.249
LatentMSA (650M)	0.639 ±0.235	6.918 ±6.340	0.664 ±0.289	0.643 ±0.298	0.638 ±0.212	8.396 ±7.271	0.655 ±0.268	0.587 ±0.282	0.788 ±0.141	4.995 ±5.186	0.836 ±0.168	0.766 ±0.195
LatentMSA (3B)	0.627 ±0.237	6.649 ±6.270	0.664 ±0.287	0.639 ±0.300	0.689 ±0.208	6.953 ±6.806	0.713 ±0.261	0.650 ±0.274	0.798 ±0.130	4.567 ±4.990	0.847 ±0.159	0.777 ±0.190
LatentMSA (15B)	0.623 ±0.241	6.901 ±6.538	0.651 ±0.297	0.631 ±0.298	0.679 ±0.207	7.030 ±6.301	0.708 ±0.251	0.645 ±0.275	0.777 ±0.140	5.412 ±5.499	0.823 ±0.172	0.749 ±0.207
LatentMSA(E2E)	**0.697** ±0.216	**5.197** ±5.667	**0.747** ±0.267	**0.726** ±0.272	**0.722** ±0.180	**5.900** ±5.689	**0.750** ±0.235	**0.692** ±0.252	**0.826** ±0.109	**3.538** ±3.977	**0.886** ±0.128	**0.827** ±0.152

Table 6: Per-target best-score comparison on CASP15, grouped by concat MSA depth, $d_{\mathrm{UniRef100}} + d_{\mathrm{ColabFoldDB}}$. For each target and method, we retain the best prediction across the evaluated seeds and samples. Values are mean ± sample standard deviation (SD) across targets within each bucket. Arrows indicate metric direction; best and second-best displayed results are bolded and underlined, respectively.

Method	[1,100), N=10				[100,500), N=8				[500,16k), N=47			
	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑
ESMFold	0.444 ±0.282	17.453 ±11.471	0.425 ±0.332	0.374 ±0.367	**0.683** ±0.226	7.759 ±7.256	**0.548** ±0.354	**0.513** ±0.353	**0.774** ±0.103	12.563 ±18.604	**0.605** ±0.346	**0.548** ±0.328
SimpleFold	0.432 ±0.248	16.027 ±9.723	0.415 ±0.304	0.345 ±0.334	0.617 ±0.203	**7.384** ±5.641	0.518 ±0.309	0.476 ±0.311	0.758 ±0.126	13.358 ±20.381	0.597 ±0.340	0.540 ±0.331
PX(NO MSA)	0.377 ±0.205	16.387 ±8.765	0.354 ±0.239	0.270 ±0.275	0.497 ±0.197	11.387 ±5.550	0.334 ±0.228	0.294 ±0.223	0.434 ±0.155	20.108 ±18.953	0.336 ±0.221	0.272 ±0.262
EvoGen	0.379 ±0.213	**13.701** ±6.979	0.365 ±0.265	0.288 ±0.295	0.511 ±0.203	9.871 ±5.568	0.381 ±0.259	0.333 ±0.273	0.389 ±0.146	20.269 ±18.498	0.279 ±0.199	0.236 ±0.231
EvoDiff	0.353 ±0.206	17.044 ±8.179	0.316 ±0.242	0.244 ±0.275	0.514 ±0.195	10.532 ±5.532	0.362 ±0.220	0.314 ±0.224	0.429 ±0.162	20.788 ±19.491	0.333 ±0.227	0.271 ±0.272
MSAGPT	0.363 ±0.205	16.024 ±8.774	0.339 ±0.247	0.257 ±0.277	0.525 ±0.192	10.541 ±5.246	0.394 ±0.262	0.347 ±0.270	0.512 ±0.180	18.578 ±19.852	0.406 ±0.263	0.338 ±0.289
LatentMSA(3B)	0.445 ±0.278	15.613 ±10.097	0.427 ±0.330	0.356 ±0.371	0.614 ±0.212	7.546 ±5.656	0.500 ±0.311	0.451 ±0.320	0.733 ±0.118	12.612 ±18.907	0.589 ±0.338	0.523 ±0.325
LatentMSA(E2E)	**0.446** ±0.260	15.424 ±9.610	**0.449** ±0.311	**0.376** ±0.347	0.628 ±0.199	7.739 ±5.487	0.512 ±0.313	0.474 ±0.313	0.752 ±0.116	**12.474** ±19.048	0.605 ±0.345	0.548 ±0.330

Table 7: Per-PDB ranking-score Top-1 comparison grouped by MSA pairing depth on CASP15. For each (PDB, method), we select the highest-ranking prediction across all seeds and samples, then average its metrics within each depth bucket. ESMFold and SimpleFold are omitted because their scoring functions differ from the Protenix-based methods, and we do not optimize or cross-calibrate ranking functions. Arrows indicate metric direction; the best method in each (bucket, metric) is bolded.

Method	[1,100), N=10				[100,500), N=8				[500,16k), N=47			
	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑	LDDT↑	RMSD↓	TM-score↑	GDT_TS↑
PX(NO MSA)	0.356 ±0.207	19.884 ±9.621	0.301 ±0.235	0.233 ±0.275	0.455 ±0.183	13.904 ±6.034	0.283 ±0.195	0.253 ±0.199	0.406 ±0.150	23.946 ±21.844	0.293 ±0.209	0.239 ±0.242
EvoGen	0.348 ±0.216	**16.717** ±8.022	0.306 ±0.260	0.248 ±0.294	0.478 ±0.198	12.406 ±5.184	0.325 ±0.234	0.287 ±0.249	0.355 ±0.139	24.524 ±22.056	0.239 ±0.187	0.202 ±0.210
EvoDiff	0.335 ±0.210	20.670 ±10.001	0.278 ±0.245	0.218 ±0.279	0.490 ±0.195	12.698 ±5.863	0.325 ±0.227	0.283 ±0.223	0.399 ±0.151	24.326 ±21.322	0.294 ±0.218	0.238 ±0.251
MSAGPT	0.345 ±0.198	18.875 ±9.924	0.293 ±0.224	0.234 ±0.264	0.488 ±0.184	11.832 ±6.054	0.356 ±0.263	0.315 ±0.262	0.489 ±0.182	21.322 ±21.779	0.378 ±0.261	0.317 ±0.281
Protenix + LatentMSA	0.416 ±0.266	17.930 ±11.327	0.385 ±0.329	0.328 ±0.365	0.566 ±0.220	10.753 ±6.810	0.431 ±0.321	0.391 ±0.325	0.705 ±0.120	15.656 ±22.417	0.547 ±0.326	0.482 ±0.308
LatentMSA(E2E)	**0.428** ±0.249	16.734 ±10.686	**0.414** ±0.305	**0.345** ±0.329	**0.607** ±0.190	**9.603** ±6.508	**0.483** ±0.310	**0.445** ±0.309	**0.725** ±0.124	**15.496** ±22.600	**0.567** ±0.337	**0.507** ±0.316

B Additional Related Work

Retrieval-Based Protein Structure Prediction. A central challenge in protein structure prediction is extracting useful information from large evolutionary and structural databases. Classical retrieval-based methods, including homology modeling and threading, search structural databases such as the PDB for homologous templates (Chothia & Lesk, 1986; Zhang, 2008). HHpred introduced HMM–HMM alignment for remote homolog detection (Söding, 2004), while I-TASSER combines threading with de novo modeling (Yang et al., 2015). These approaches can be effective when close templates are available, but their accuracy degrades when homologous templates are sparse or unreliable.

Deep learning methods have transformed how retrieved information is integrated. AlphaFold2 incorporates MSAs and structural templates as explicit conditioning signals for end-to-end structure prediction (Jumper et al., 2021). RoseTTAFold uses a three-track architecture to exchange information among sequence features, distance maps, and 3D coordinates (Baek et al., 2021). ColabFold further accelerates MSA retrieval through MMseqs2, making large-scale homology search more practical (Mirdita et al., 2022). In parallel, MSA-free predictors such as ESMFold (Lin et al., 2023) and OmegaFold (Wu et al., 2022) reduce dependence on explicit homology retrieval by leveraging protein language model representations. More recently, structure-based retrieval tools such as Foldseek (van Kempen et al., 2024) have enabled retrieval over structural databases, providing complementary geometric priors for downstream prediction. These developments highlight the continued importance of retrieval while motivating complementary approaches, such as LatentFold, that perform no target-specific homology retrieval at inference time.