# Search, Edit, and Fold: LLM-Guided MSA Optimization for Protein Conformation Prediction

**Yu Pei**[12] **Jiangtao Feng**[1] **Hao Wang**[12] **Dongyu Xue**[1] **Keyue Qiu**[12]
**Hao Zhou**[12*] **Wei-Ying Ma**[12]
[1] Institute of AI Industry Research (AIR), Tsinghua University
[2] Department of Computer Science and Technology, Tsinghua University
yupei.wp@gmail.com
{zhouhao,maweiying}@air.tsinghua.edu.cn

## Abstract

Accurately identifying alternative protein conformations remains a fundamental challenge, particularly when functionally relevant states are encoded by sparse evolutionary signals within large multiple sequence alignments (MSAs). In this work, we reformulate protein conformation prediction as a combinatorial search problem in MSA space, shifting the focus from structure divergence to evolutionary information discovery. We introduce **MSA-Evolver**, an optimization framework that enables LLM with direct manipulation and iterative exploration of MSAs. Specifically, MSA-Evolver introduces a unified action space for MSA editing together with a feedback-guided multi-step reasoning strategy that allows the language model to progressively explore, evaluate, and refine candidate MSAs based on historical search trajectories. Our framework efficiently identifies informative sub-MSAs under limited folding budgets and substantially improves the prediction accuracy of alternative conformations, including open-closed, inward-outward, apo-holo, fold-switching, and intrinsically disordered proteins.

## 1 Introduction

Proteins perform their biological functions including transcription [Tan and Richmond, 1990, Roberts and Green, 1994], metabolism [Li and Poulos, 1999, Rathmell et al., 2003] and intracellular communication [Schneider and Simons, 2013] by transitioning between multiple conformational states. Thus, identification of these alternative conformations is essential for understanding and discovery of a biological mechanism. Despite the success of AlphaFold [Jumper et al., 2021, Abramson et al., 2024] series in modeling static structures and protein interactions, they often underperform in identifying remote alternative conformations compared to computationally intensive molecular dynamics (MD) simulations. This limitation is closely related to their dependence on multiple sequence alignments (MSAs), which serve as the primary source of evolutionary constraints for structure inference [Lazou et al., 2024].

The challenges of utilizing MSA information in protein conformation prediction lie in the imbalance of MSA quantity and quality at evolutionary level. Modern retrieval pipelines such as MMseqs2 [Steinegger and Söding, 2017] and ColabFold [Mirdita et al., 2022] produce increasingly large MSAs, however, evolutionary homology signals that encode conformational switching are typically sparse and subtle. This imbalance suggests that the key challenge is not merely generating diverse structures, but also identifying and isolating informative sub-MSAs that preserve these functionally critical signals.

*Correspondence to Hao Zhou (zhouhao@air.tsinghua.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

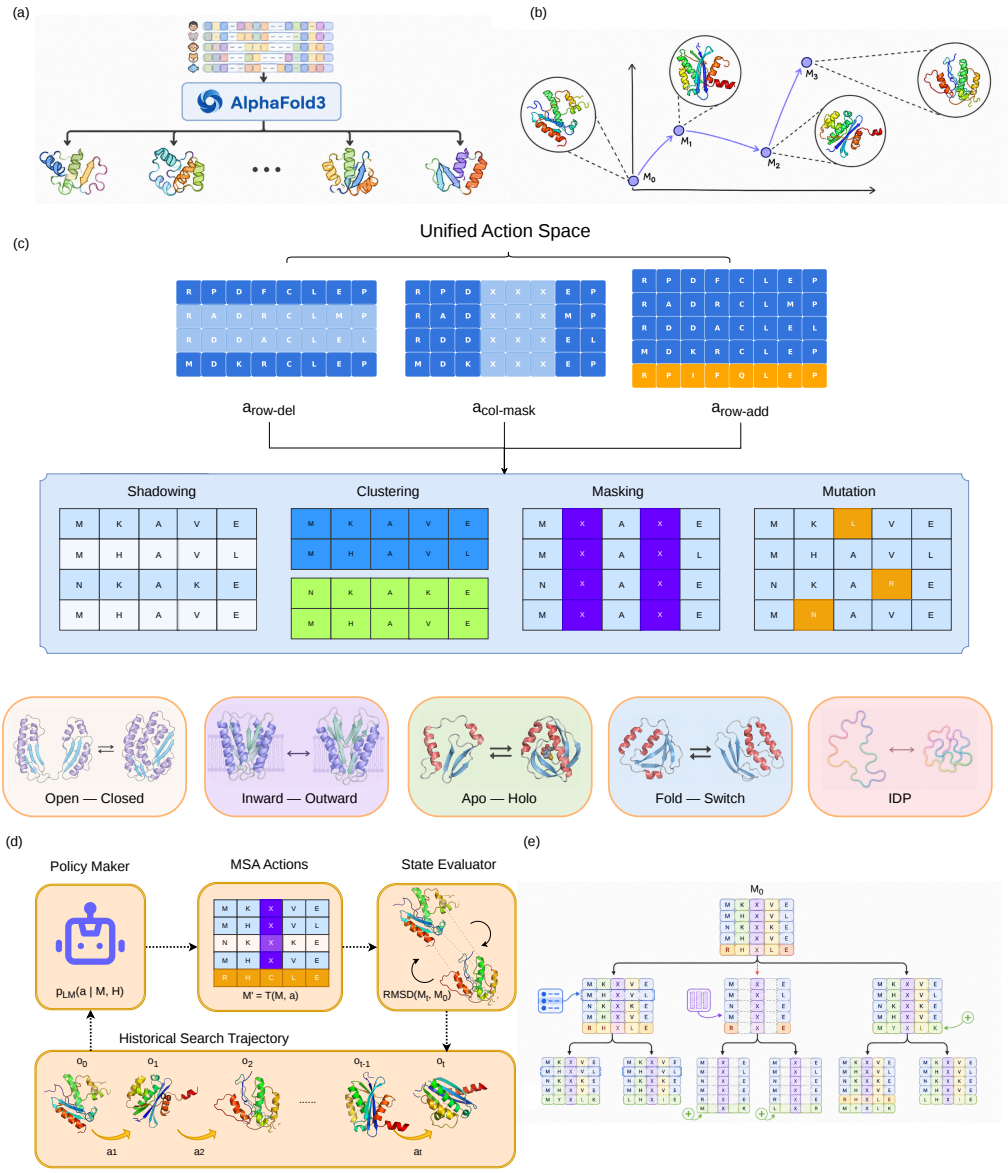



Figure 1: **(a)** Protein conformation prediction has been considered a task of structure ensemble. **(b)** We reformulate protein conformation prediction as a combinatorial searching problem in MSA Space. **(c)** Unified action space is construct for LLM to systematically operate MSA. **(d)** A framework of LLM-driven multi-step reasoning and iterative search. **(e)** Searching strategy in MSA Space balancing exploration and efficiency.

To extract specialized sub-MSAs, previous studies typically rely on biologically inspired and manually designed MSA manipulation strategies, such as shadowing [Del Alamo et al., 2021], filtering [Zhang et al., 2026], masking [Kalakoti and Wallner, 2025, 2026], clustering [Lazou et al., 2024], and introducing sequence mutations [Stein and Mchaourab, 2022]. These approaches demonstrate that modification of MSA can substantially influence the conformational preferences of folding models. However, these pipelines are predominantly rule-based and heavily rely on biological expert intervention to design and select MSA operations, limiting their scalability and automation.

Motivated by these limitations, we reformulate protein conformation prediction as a combinatorial search problem in MSA space, where each state corresponds to a modified MSA and each transition

represents a unified operation over sequences or residues. From this perspective, the objective is no longer to manually engineer specialized MSA heuristics but to automatically discover informative combinations of MSA editing operations that can effectively modulate folding behavior. This reformulation naturally motivates the introduction of large language models (LLMs). Recent advances have shown that LLMs possess strong capabilities in structured reasoning [Patil and Jadon, 2025], iterative decision-making [Cheng et al., 2025], and navigating complex search spaces [Xiong et al., 2024, Xi et al., 2025]. More importantly, LLMs provide a more meta-level paradigm for protein conformation prediction: instead of directly designing folding algorithms and models, LLMs could act as intelligent controllers that adaptively optimize the inputs of powerful folding models through iterative MSA exploration.

However, directly applying LLMs to MSA manipulation remains challenging. First, MSAs can be extremely large, exceeding the effective context length of LLMs [Liu et al., 2025, Bai et al., 2025]. Second, general-purpose LLMs are not trained to model protein evolutionary landscapes or sequence–structure coupling, thus lack the ability to retrieve homolog combinations or estimate the structural effects of MSA operations [Heinzinger et al., 2024, Chen et al., 2024].

To address the obstacles above, we propose **MSA-Evolver, an LLM-driven search framework for MSA optimization**. As shown in Figure 1, instead of passively encoding the entire MSA, we introduce a unified action space for MSA manipulation, which defines the transition operators of the MSA search process. To enable scalable exploration, we introduce a feedback-oriented multi-step reasoning mechanism, enabling the LLM to iteratively explore, evaluate, and refine candidate MSAs through a structured search process with branching and backtracking. Empirically, our method consistently improves the prediction of diverse conformational transitions, including **open-closed, inward-outward, apo-holo, fold-switching and intrinsically disordered proteins** with limited folding budget. The contributions of this work are summarized as follows:

- We reformulate protein conformation prediction as a combinatorial search problem in MSA space, shifting the focus from structure generation to evolutionary signal retrieval.
- We propose a unified MSA action space with minimal yet expressive operations, enabling flexible and systematic exploration of the MSA space.
- We develop a feedback-driven multi-step reasoning framework that allows LLMs to perform iterative search with evaluation and refinement at test time.

## 2 MSA-Evolver

Aiming to explore the high-dimensional MSA space for alternative protein conformations, MSA-Evolver empowers an LLM to reason over historical search trajectories and iteratively propose MSA modifications, thereby enabling selective conformational exploration. We first formulate conformation discovery as an MSA exploration problem, defining the search objective and state space over which the LLM operates. Next, we construct a unified MSA editing action space, including row shadowing, row addition, and column masking, which enables the LLM to perform structured state transitions through diverse evolutionary perturbations. Finally, we model the entire procedure as a multi-step decision-making process, allowing the LLM to leverage trajectory-level context for progressive and context-aware exploration in the MSA space.

### 2.1 Problem Setup: Conformation Searching

Formulating MSA manipulation as a combinatorial search problem rather than structure ensembling, our aim is to find remote conformations in finite attempts. Let a multiple sequence alignment (MSA) be represented as a matrix:

$$M \in \Sigma^{N \times L} \quad \Sigma = \{\mathrm{AA}\} \cup \{-, X\} \tag{1}$$

where AA stands for 20 amino acid, $-$ stands for gap, $X$ means uncertain residue which is also masked residue, $N$ is depth of searched MSA, $L$ means protein sequence length.

Therefore, the MSA search space is defined as $\mathcal{M} = \bigcup_{N=1}^{\infty} \Sigma^{N \times L}$ Given an initial MSA $M_0$, our goal is to search within the MSA space $\mathcal{M}$ to identify alternative MSAs that induce different structural

Table 1: Unified formulation of MSA editing operations.

| MSA Actions | Row Shadowing | Row Addition | Column Masking |
|---|---|---|---|
| ***Symbol*** | $a_{row-del}$ | $a_{row-add}$ | $a_{col-mask}$ |
| ***Transformation*** | $M' = M[\mathcal{I}, :]$ | $M' = \begin{bmatrix} M \\ \widetilde{M} \end{bmatrix}$ | $M'_{i,j} = \begin{cases} X, & j \in \mathcal{J} \\ M_{i,j}, & \text{otherwise} \end{cases}$ |
| ***Shape*** | $\Sigma^{N \times L} \to \Sigma^{N' \times L}$ | $\Sigma^{N \times L} \to \Sigma^{(N+\Delta N) \times L}$ | $\Sigma^{N \times L} \to \Sigma^{N \times L}$ |
| ***Deterministic*** | $\mathcal{I} \subset \{1, \ldots, N\}$ | $\widetilde{M} \sim p_{\text{PLM}}(\cdot \mid M)$ | $\mathcal{J} \subset \{1, \ldots, L\}$ |

conformations:

$$M^* = \arg\max_{M \in \mathcal{M}} V(M) \tag{2}$$

$V(M)$ is an evaluation function that measures the structural difference induced by MSA $M$, which will be integrated later into a partially observable Markov decision process(POMDP) process. $M^*$ means the potential optimal point.

### 2.2 Action Space Construction

To solve the combinatorial puzzle in the MSA space, we construct a unified action space together with a policy-driven generation mechanism. Given the current state $s = M$, the language model acts as a stochastic policy that proposes candidate actions conditioned not only on the current MSA $M$ but also on historical information $\mathcal{H}$:

$$a^{(i)} \sim p_\theta(a \mid M, \mathcal{H}), \quad i = 1, \ldots, k \tag{3}$$

where $\mathcal{H}$ denotes the accumulated context, including explored actions and their evaluation scores. Each action $a$ induces a transition in the MSA space $M' = T(M, a)$ where $T(\cdot)$ is a transition operator defined over MSA matrices. Therefore, the thought generator $G$ can be formulated as a policy-guided proposal operator:

$$G(p_\theta, M, k, \mathcal{H}) = \{T(M, a^{(i)})\}_{i=1}^{k} \tag{4}$$

Where $k$ nodes are expanded from the current state $M$. As shown in Table 1, we support 3 types of actions as basis, expecting to cover the MSA space with their compositions:

$$a = a_{\text{col-mask}} \circ a_{\text{row-add}} \circ a_{\text{row-del}} \tag{5}$$

**(1) Row Shadowing** $a_{row-del}$ **:** Row shadowing action operates MSA by selecting a subset of informative rows. Covering MSA clustering as a particular case, this operation can be interpreted as finding specific evolutionary couplings and removing low-quality homologous sequences, thereby sharpening the signal of evolutionary constraints. By dynamically controlling sequence diversity and depth, row shadowing allows the search process to focus on more relevant evolutionary patterns and prevents the model from being dominated by noisy or biased alignments.

**(2) Row Addition** $a_{row-add}$ **:** Row addition action operates MSA by replenishing evolutionary information inside protein language models. Conditioned on the existing alignment, this operation expands the original MSA space by introducing new sequences beyond the observed data. This causes direct distributional shifts, thus being effective when exploring underrepresented or missing regions of the MSA landscape.

**(3) Column Masking and Re-Masking** $a_{col-mask}$ **:** Column masking and re-masking action operate MSA by breaking existing co-evolutionary relations between residues. Replacing selected positions with an unknown token $X$, this operation weakens specific positional constraints, forcing the structure prediction model to reconsider structural dependencies under partial information, thus enclose more possibilities and potential for alternative conformations.

**Coverage of Action Space** Iterative search is conducted each step appling a composed action $a_i$, the reachable MSA space induced by $G$ after $t$ steps is:

$$\mathcal{M}_t^{\text{comp}}(M_0) = \{M_t \mid M_t = T(\cdot, a_t) \circ \cdots \circ T(M_0, a_1),\ a_i \sim p_\theta(\cdot \mid M_{i-1})\} \tag{6}$$

Compared with prior methods that rely on a single type of MSA manipulation, the proposed compositional action space subsumes a significantly broader set of reachable MSA states.

**Proposition 1** *The composed action space subsumes any single-action reachable space:*

$$\forall \textit{ single action type } \alpha, \quad \mathcal{M}_t^{\alpha}(M_0) \subseteq \mathcal{M}_t^{\textit{comp}}(M_0) \tag{7}$$

where $\mathcal{M}_t^{\text{single}}(M_0)$ denote the reachable space under a single type of operation, including row shadowing , row clustering, column masking, and single residue mutations. Proofs and explanations found in Appendix A. Importantly, this unified formulation not only enlarges the searchable MSA space but also provides a structured interface for the LLM to perform iterative reasoning and trajectory-level planning, which forms the foundation of the multi-step exploration framework introduced next.

### 2.3 Multi-step Context Exploration

With the unified MSA action space defined above, MSA optimization naturally becomes a sequential decision-making problem where we allow the model to iteratively compose actions, forming a trajectory of edits over the MSA. This formulation enables structured exploration of the exponentially large MSA space and provides a principled way to balance exploration and exploitation through multi-step reasoning.

**Conformation Search as a POMDP** A key challenge through MSA exploration is that the quality of an MSA is not directly observable. The usefulness of an edited MSA can only be assessed indirectly through the structures predicted by a folding model. To capture this partial observability, we formulate MSA optimization for protein conformation prediction as a POMDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, \Omega = p_F, R, \gamma \approx 0, b_0 = \{M_0\} \rangle$.

where:

- **State** $\mathcal{S}$: MSA states $s_t = M_t$, with initial state $s_0 = M_0$.
- **Action** $\mathcal{A}$: Composed MSA editing operations defined in Section 2.2.
- **Observation** $\mathcal{O}$: Predicted structures $y_{t+1} \sim p_F(\cdot \mid s_{t+1})$ from the folding model.
- **Transition Function** $T$: $s_{t+1} = T(s_t, a_t)$.
- **Reward** $R$: $R(s_t) = \frac{1}{|\mathcal{B}_t|} \sum_{y' \in \mathcal{B}_t} \text{RMSD}(y_t, y')$
- **Belief** $b_t$: A set of anchor structures $\mathcal{B}_t$.

Under this formulation, the LLM acts as a policy:

$$\pi = p_\theta(a_t \mid s_t, \mathcal{H}_t), \quad \mathcal{H}_t = \mathcal{H}_{t-1} \cup \{(a_t, R(s_t))\} \tag{8}$$

where $\mathcal{H}_t$ aggregates historical actions and their outcomes, enabling context-aware decision making, and proxy reward metric $R(s_t)$ encourages structural diversity and promotes exploration of alternative conformations. At each step, the language model must balance proposing diverse MSA edits to discover unseen conformations and refining promising MSA regions that already yield high-quality structures. The diversity-based reward and historical context together guide this trade-off, encouraging the model to avoid premature convergence while still focusing on high-potential regions of the search space.

**Multi-step Reasoning through POMDP** The formulation above allows the LLM to perform multi-step reasoning by explicitly conditioning on past actions and rewards. To approximate the latent state of the POMDP, we maintain a belief state represented by a set of structurally diverse anchors $b_t = \mathcal{B}_t$. A newly predicted structure is added to the belief set if it is sufficiently different:

$$y_t \in \mathcal{B}_t \quad \text{if} \quad \min_{y' \in \mathcal{B}_{t-1}} \text{RMSD}(y_t, y') > \delta \tag{9}$$

Table 2: Mean best-matching TM-score between predicted structures and target conformations across four types of conformationally changing proteins. Scores are reported with respect to the preferred and alternative conformations. The best result in each column is shown in **bold**.

| Method | Open–Close | | In–Out | | Apo–Holo | | Fold–Switch | |
|---|---|---|---|---|---|---|---|---|
| | Pref. | Alter. | Pref. | Alter. | Pref. | Alter. | Pref. | Alter. |
| AlphaFold2 | 0.911 | 0.806 | 0.827 | 0.544 | 0.868 | 0.757 | 0.774 | 0.636 |
| AlphaFold3 | 0.926 | 0.800 | 0.828 | 0.532 | 0.865 | 0.751 | 0.780 | 0.620 |
| *Structure-based* | | | | | | | | |
| ESM-Diff | 0.874 | 0.793 | 0.796 | 0.544 | 0.856 | 0.755 | 0.690 | 0.569 |
| ProtTeX | 0.469 | 0.437 | 0.458 | 0.390 | 0.781 | 0.692 | 0.610 | 0.490 |
| AlphaFlow | 0.914 | 0.811 | 0.828 | 0.580 | 0.874 | 0.776 | 0.780 | 0.636 |
| *MSA-based* | | | | | | | | |
| AF-Cluster | 0.869 | 0.807 | 0.724 | 0.588 | 0.854 | 0.772 | 0.717 | 0.601 |
| MSA-Subs. | 0.920 | 0.843 | 0.839 | 0.642 | 0.881 | 0.783 | 0.784 | 0.643 |
| AFSample2 | 0.933 | 0.847 | **0.862** | **0.697** | 0.880 | 0.778 | 0.792 | 0.641 |
| AFSample3 | 0.922 | 0.850 | 0.829 | 0.668 | 0.879 | 0.784 | 0.771 | 0.630 |
| **MSA-Evolver** | **0.940** | **0.854** | 0.861 | 0.665 | **0.888** | **0.797** | **0.852** | **0.683** |

This mechanism explicitly tracks explored conformational modes and provides a memory of the search process. To further stabilize exploration, we maintain a beam of candidate MSA states:

$$S_t = \mathrm{Top}_k \left( \bigcup_{s \in S_{t-1}} G(p_\theta, s, k, \mathcal{H}_{t-1}) \right) \tag{10}$$

This allows parallel exploration of multiple trajectories while retaining the most promising candidates. The optimal state is then selected as:

$$s_t^* = \arg\max_{s \in S_t} R(s) \tag{11}$$

Overall, by combining a unified action space, belief-guided exploration, and LLM-driven policy reasoning, MSA-Evolver enables efficient navigation of the MSA space and achieves a principled balance between exploration and exploitation.

## 3 Experiments

The effectiveness of MSA-Evolver is widely confirmed across various types of proteins. In this section, we first introduce an integrated evaluation dataset and our main experimental result on conformation changing pairs. Then followed by conformation prediction of intrinsically disordered proteins(IDPs) and case study of multiple conformation protein: Bovine pancreatic trypsin inhibitor(BPTI). Ablation studies also demonstrate the contribution of MSA-Evolver's unified action space and exploration strategy. For more

Table 3: Best matching RMSD compared to 5 kinetic BPTI conformations. The Conf.3 column is highlighted as the most challenging and key conformation. Best performing result is **bolded** and second-best value is underlined.

| RMSD(Å)↓ | Conf.1 | Conf.2 | Conf.3 | Conf.4 | Conf.5 | Avg. |
|---|---|---|---|---|---|---|
| *Structure-based* | | | | | | |
| ProtTeX | 1.74 | 2.59 | 2.33 | 1.97 | 1.93 | 2.11 |
| ESMDiff | 0.90 | 1.58 | 2.03 | **1.28** | 1.05 | 1.37 |
| *MSA-based* | | | | | | |
| AF_cluster | 0.92 | 1.46 | 1.78 | 1.35 | 0.96 | 1.29 |
| MSA-Subs | 0.91 | 1.60 | 2.16 | 1.57 | 0.97 | 1.44 |
| AFSample2 | **0.87** | 1.61 | 2.15 | 1.48 | **0.91** | 1.40 |
| AFSample3 | 0.92 | 1.57 | 2.16 | 1.62 | 0.97 | 1.45 |
| **MSA-Evolver** | 0.89 | **1.31** | **1.76** | 1.54 | **0.91** | **1.28** |

experimental details, please refer to Appendix F.

### 3.1 Main Results: Prediction of Conformation Changing Pairs

**Evaluation Datasets** To comprehensively evaluate the ability of MSA-Evolver to predict diverse conformations beyond AlphaFold3 [Abramson et al., 2024], we construct a hybrid benchmark consisting of 227 proteins with experimentally verified conformational changes. These proteins cover four representative categories of biological processes: **(1) 45 Open/Closed** state protein pairs following [Kalakoti and Wallner, 2025], which is the union of dataset OC23 and OC23 $>$ 85. **(2) 16 Inward/Outward** transporter proteins [Kalakoti and Wallner, 2025], which mainly involve membrane transport processes accompanied by large-scale conformational rearrangements. **(3) 90 Apo/Holo** pairs with ligand-induced conformational changes following [Jing et al., 2023]. (4) 76 Fold-Switching proteins following [Chakravarty and Porter, 2022], which are collected from PDB [Burley et al., 2017] but AlphaFold2 [Jumper et al., 2021] fails to capture both state. All protein structures listed above are associated with unique PDB identifiers. The detailed lists of protein targets are provided in Appendix E.

**Baselines** To provide a comprehensive and fair comparison for evaluating the effectiveness of MSA-Evolver, we consider two categories of baseline methods.

- **AlphaFold-series.** We include AlphaFold2 [Jumper et al., 2021] and AlphaFold3 [Abramson et al., 2024, Team et al., 2025] as structure-prediction baselines.
- **Structure-based methods.** Parameterized generative models have recently demonstrated strong capabilities in exploring the structural sample space of proteins. **AlphaFlow** [Jing et al., 2024] applies flow matching to generate diverse protein conformations. **ProtTeX** [Ma et al., 2025] is a hybrid language model designed to model the joint sequence–structure–function space of proteins, trained on additional self-constructed datasets. **ESM-Diff** [Lu et al., 2024] integrates the evolutionary information captured by masked language models with a structure decoder based on masked diffusion.
- **MSA-based methods.** These approaches build upon the inference pipeline of AlphaFold3 [Abramson et al., 2024] and focus on manipulating retrieved MSAs, whose variations can significantly influence the resulting structural conformations. We include **AF-cluster** [Wayment-Steele et al., 2024], a representative method that discovers diverse conformations through clustering analysis of MSA representations; **MSA-Subsampling** (MSA-Subs.), an approach that identifies alternative conformations by shadowing MSAs; and **AFSample2** [Kalakoti and Wallner, 2025] and **AFSample3** [Kalakoti and Wallner, 2026], a series of methods that encourage conformational diversity by masking MSA columns to disrupt original co-evolutionary signals.

**Metrics** As molecular dynamics(MD) data are not included in our setting, we primarily adopt Template Modeling Score(TM-score) [Zhang and Skolnick, 2005] and Root-Mean-Square Deviation(RMSD) as evaluation metrics to assess the structural accuracy of MSA-Evolver and competing baselines. To mitigate the potential bias introduced by the default structural preference of AlphaFold3 [Abramson et al., 2024] in prediction two conformations, we compute TM-scores separately with respect to the AlphaFold3 preferred conformation and the alternative conformation, following the evaluation protocol in [Kalakoti and Wallner, 2026]. In this setting, the TM-score with respect to the alternative conformation is particularly important, as it better reflects the ability of a method to capture non-dominant structural states.

As shown in Table 2, MSA-Evolver achieves best performance on most evaluation settings, outperforming both MSA-based and structure-based baselines in predicting the preferred as well as the alternative conformations. This advantage remains stable across all four functional protein subsets, demonstrating the strong generalization ability of MSA-Evolver across diverse conformational switching states. Beyond global metrics, qualitative analysis of the 3D structures further reveals that MSA-Evolver produces conformations with better local agreement to the ground truth. In particular, key secondary structure elements—especially $\alpha$ helices—are more accurately recovered. As highlighted in the red-box regions in Figure 2, our predictions exhibit clearer and more precise helix

orientations compared to competing methods, which often suffer from distortion or misalignment in these regions.

### 3.2 Prediction of intrinsically disordered proteins

Next, MSA-Evolver is extended to predict changing ensembles of intrinsically disordered proteins, which are believed to be abundant in nature, but without adopting a single well-defined structure [Oldfield and Dunker, 2014, Trivedi and Nagarajaram, 2022].

**Metrics** Since IDPs do not have a fixed or stable tertiary structure under normal conditions, alignment-based metrics such as TM-score or RMSD are not suitable. Instead, following [Lu et al., 2024], mean absolute error(MAE) is calculated to describe the deviation between generated ensembles and experimental structural data. Three properties are calculated and compared between in-silico predicted structure and structure ensembles in dataset: pairwise distance, radius of gyration and contact map. Detailed information of IDP datasets and metrics are included in Appendix E.2, F.4.

Table 4: Mean MAE (mean absolute error) across IDPs between 3 properties: Pairwise Distance (PwD.), Radius of gyration (Rg.) and Contact map (Cp.). The top-performing method is **bolded** and the second-best result is underlined.

| Method | PwD.↓ | Rg.↓ | Cp.↓ |
|---|---|---|---|
| *Structure-based* | | | |
| ProtTeX | 9.061 | 4.280 | 0.173 |
| ESM-Diff | 8.048 | 4.231 | 0.151 |
| *MSA-based* | | | |
| AF-Cluster | 8.769 | 4.955 | 0.158 |
| MSA-Subs. | 7.649 | **3.915** | 0.132 |
| AFSample2 | 7.603 | 4.200 | **0.127** |
| AFSample3 | 7.782 | 4.267 | 0.131 |
| **MSA-Evolver** | **7.361** | 4.003 | **0.127** |

From Table 4, MSA-Evolver consistently achieves competitive performance across all ensemble-level statistics, showing its prospect in exploring conformation ensembles while reserving experimental-aligned properties.

### 3.3 Case Study: Multi-Conformational BPTI

Bovine pancreatic trypsin inhibitor (BPTI) serves as a representative landscape for MSA-Evolver to capture multiple conformational states. Comprising 5 established structures [Huber et al., 1974, Brooks and Karplus, 1983], common methods failed on the recognition of remote cluster 3. As shown in Table 3, MSA-Evolver successfully covers most of the known conformational clusters and achieves superior performance. In particular, MSA-Evolver demonstrates an advantage in recovering the difficult Cluster 3 state, highlighting its strong capability in exploring diverse protein conformational spaces.

### 3.4 Ablation Study

To validate the effectiveness of the unified action space and the proposed search strategy, we conduct ablation studies from two perspectives. First, as shown in Table 5, progressively restricting MSA editing operations leads to consistent performance degradation, indicating that each operation contributes to prediction quality. The combination of a minimal yet expressive action set enables MSA-Evolver to expand the searchable MSA space and improve exploration during folding. Second, we compare MSA-Evolver with standard search algorithms, including Depth-First Search (DFS) and Breadth-First Search (BFS), applied to the same action space without reward guidance or belief-state reasoning. As shown in Table 7, both methods perform worse than MSA-Evolver, demonstrating that naive traversal is insufficient. In contrast, MSA-Evolver leverages feedback-driven, multi-step reasoning to guide exploration toward more promising regions.

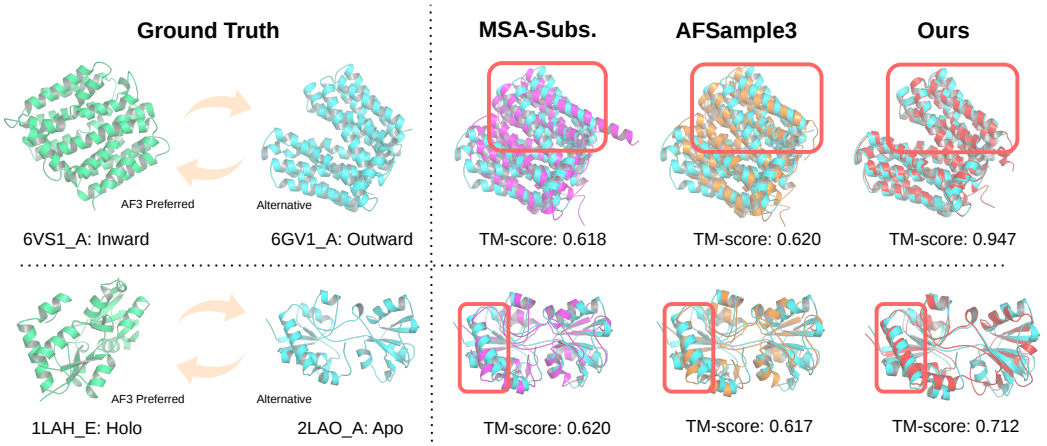



Figure 2: Case Study of conformation changing pairs: transporter MdfA(upper) and Salmonella typhimurium LAO binding protein(lower). Detailed information refer to Appendix F.5.

# 4 Related Works

## 4.1 MD simulation and MD-data based Protein Dynamics

Molecular dynamics (MD) simulations provide a physically grounded approach to exploring protein conformational landscapes by explicitly modeling atomic interactions over time. However, their practical application is fundamentally limited by the high computational cost and the difficulty of adequately sampling high-dimensional energy landscapes. Recent works attempt to amortize this cost by training deep generative models on MD trajectories, such as AlphaPPImd [Wang et al., 2024a], idpGAN [Janson et al., 2023] and AI$^2$BMD [Wang et al., 2024b]. While these methods can reproduce physically plausible motions, they inherit several intrinsic limitations from MD data. First, MD simulations are biased toward low-energy states and may fail to capture rare but functionally important conformations due to insufficient sampling. Second, the generated conformational distributions are tied to the simulated trajectories of individual sequences, lacking the evolutionary diversity observed across homologs.

## 4.2 Structural generative model in Conformation Prediction

Recent advances in generative modeling, including diffusion and flow-matching approaches, have enabled the direct sampling of diverse protein conformations. Methods such as EigenFold [Jing et al., 2023], ConfDiff [Wang et al., 2024c] incorporate physical priors or energy-based constraints to improve the realism of generated structures. Other works, including BioEmu [Lewis et al., 2025] and DiG [Zheng et al., 2024], further leverage large-scale or specialized datasets to model broader protein dynamics. Despite these advances, such approaches typically require substantial additional training and rely on carefully designed conditioning signals, such as secondary structures, to guide generation.

# 5 Conclusion

We reformulate protein conformation prediction as a combinatorial search problem in MSA space and introduce a unified, minimal action space for systematic exploration. Building on this, we propose a feedback-driven multi-step reasoning framework that enables large language models to act as adaptive search policies at test time. Experiments across diverse protein categories and representative cases show that MSA-Evolver consistently improves the discovery of alternative conformations with only a few folding samples, suggesting a promising direction for integrating LLM-based structured search and reasoning into biological sequence modeling and structure prediction.

## 6 Broader and Social Impacts

As the study focuses on the technique of protein conformation prediction, its impact may lie in biological scientific discovery, especially alternative protein conformations. Furthermore, the study highly reduce the budgets for finding valid protein conformations, which would possibly boost the pipeline of drug discovery and synthetic biology production.

## 7 Acknowledgments

This work is supported by the Natural Science Foundation of China (Grant No. 62376133) and sponsored by Beijing Nova Program (20240484682). We thank all the co-authors for their contribution and hard-work. In addition, insightful suggestions and help are valued from Yangyang Xu, Yuxuan Song, Lihao Wang, Botian Wang and Zhicheng Cai according to this research project.

## References

Song Tan and Timothy J Richmond. Dna binding-induced conformational change of the yeast transcriptional activator prtf. *Cell*, 62(2):367–377, 1990.

Stefan GE Roberts and Michael R Green. Activator-induced conformational change in general transcription factor tfiib. *Nature*, 371(6499):717–720, 1994.

Huiying Li and Thomas L Poulos. Fatty acid metabolism, conformational change, and electron transfer in cytochrome p-450bm-3. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1441(2-3):141–149, 1999.

Jeffrey C Rathmell, Casey J Fox, David R Plas, Peter S Hammerman, Ryan M Cinalli, and Craig B Thompson. Akt-directed glucose metabolism can prevent bax conformation change and promote growth factor-independent survival. *Molecular and cellular biology*, 23(20):7315–7328, 2003.

Anja Schneider and Mikael Simons. Exosomes: vesicular carriers for intercellular communication in neurodegenerative disorders. *Cell and tissue research*, 352(1):33–47, 2013.

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.

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.

Maria Lazou, Omeir Khan, Thu Nguyen, Dzmitry Padhorny, Dima Kozakov, Diane Joseph-McCarthy, and Sandor Vajda. Predicting multiple conformations of ligand binding sites in proteins suggests that alphafold2 may remember too much. *Proceedings of the National Academy of Sciences*, 121 (48):e2412719121, 2024.

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.

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.

Diego Del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling the conformational landscapes of transporters and receptors with alphafold2. *BioRxiv*, pages 2021–11, 2021.

Jing Zhang, Rongqing Yuan, Andriy Kryshtafovych, Jimin Pei, Rachael C Kretsch, R Dustin Schaeffer, Jian Zhou, Rhiju Das, Nick V Grishin, and Qian Cong. Assessment of protein complex predictions in casp16: Are we making progress? *Proteins: Structure, Function, and Bioinformatics*, 94(1): 106–130, 2026.

Yogesh Kalakoti and Björn Wallner. Afsample2 predicts multiple conformations and ensembles with alphafold2. *Communications biology*, 8(1):373, 2025.

Yogesh Kalakoti and Björn Wallner. Afsample3: Generating and selecting multiple conformational states with alphafold3. *bioRxiv*, pages 2026–01, 2026.

Richard A Stein and Hassane S Mchaourab. Speach_af: Sampling protein ensembles and conformational heterogeneity with alphafold2. *PLoS computational biology*, 18(8):e1010483, 2022.

Avinash Patil and Aryan Jadon. Advancing reasoning in large language models: Promising methods and approaches. *arXiv preprint arXiv:2502.03671*, 2025.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert Van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. *arXiv preprint arXiv:2502.15652*, 2025.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*, 17(6):4558–4577, 2024.

Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu, and Weinan Zhang. A survey of llm-based deep search agents: Paradigm, optimization, evaluation, and challenges. *arXiv preprint arXiv:2508.05668*, 2025.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, 2025.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 11 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae150. URL `https://doi.org/10.1093/nargab/lqae150`.

Dexiong Chen, Philip Hartout, Paolo Pellizzoni, Carlos Oliver, and Karsten Borgwardt. Endowing protein language models with structural knowledge. *arXiv preprint arXiv:2401.14819*, 2024.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

Devlina Chakravarty and Lauren L Porter. Alphafold2 fails to predict protein fold switching. *Protein Science*, 31(6):e4353, 2022.

Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, pages 627–641, 2017.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, et al. Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. *BioRxiv*, pages 2025–01, 2025.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

Zicheng Ma, Chuanliu Fan, Zhicong Wang, Zhenyu Chen, Xiaohan Lin, Yanheng Li, Shihao Feng, Ziqiang Cao, Jun Zhang, and Yi Qin Gao. Prottex: Structure-in-context reasoning and editing of proteins with large language models. *Journal of Chemical Information and Modeling*, 65(13): 6599–6612, 2025.

Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and Jian Tang. Structure language models for protein conformation generation. *arXiv preprint arXiv:2410.18403*, 2024.

Hannah K Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, 625(7996):832–839, 2024.

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

Christopher J Oldfield and A Keith Dunker. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual review of biochemistry*, 83(1):553–584, 2014.

Rakesh Trivedi and Hampapathalu Adimurthy Nagarajaram. Intrinsically disordered proteins: an overview. *International journal of molecular sciences*, 23(22):14050, 2022.

Robert Huber, Dietmar Kukla, Wolfram Bode, Peter Schwager, Klaus Bartels, Johann Deisenhofer, and Wolfgang Steigemann. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor: Ii. crystallographic refinement at 1.9 å resolution. *Journal of Molecular Biology*, 89(1):73–101, 1974.

Bernard Brooks and Martin Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences*, 80(21): 6571–6575, 1983.

Jianmin Wang, Xun Wang, Yanyi Chu, Chunyan Li, Xue Li, Xiangyu Meng, Yitian Fang, Kyoung Tai No, Jiashun Mao, and Xiangxiang Zeng. Exploring the conformational ensembles of protein–protein complex with transformer-based generative model. *Journal of chemical theory and computation*, 20(11):4469–4480, 2024a.

Giacomo Janson, Gilberto Valdes-Garcia, Lim Heo, and Michael Feig. Direct generation of protein conformational ensembles via machine learning. *Nature Communications*, 14(1):774, 2023.

Tong Wang, Xinheng He, Mingyu Li, Yatao Li, Ran Bi, Yusong Wang, Chaoran Cheng, Xiangzhen Shen, Jiawei Meng, He Zhang, et al. Ab initio characterization of protein molecular dynamics with ai2bmd. *Nature*, 635(8040):1019–1027, 2024b.

Yan Wang, Lihao Wang, Yuning Shen, Yiqun Wang, Huizhuo Yuan, Yue Wu, and Quanquan Gu. Protein conformation generation via force-guided se (3) diffusion models. *arXiv preprint arXiv:2403.14088*, 2024c.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761): eadv9817, 2025.

Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, 6(5):558–567, 2024.

Alex L Zhang, Tim Kraska, and Omar Khattab. Recursive language models. *arXiv preprint arXiv:2512.24601*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Jieyi Long. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023.

Shentong Mo and Miao Xin. Tree of uncertain thoughts reasoning for large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12742–12746. IEEE, 2024.

Jonathan Silva, Qin Ma, Jordi Cabot, Pierre Kelsen, and Henderik A Proper. Application of the tree-of-thoughts framework to llm-enabled domain modeling. In *International Conference on Conceptual Modeling*, pages 94–111. Springer, 2024.

Juntong Wu, Zijing Liu, He Cao, Li Hao, Bin Feng, Zishan Shu, Ke Yu, Li Yuan, and Yu Li. Rethinking text-based protein understanding: Retrieval or llm? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23737–23757, 2025.

Jingjing Gong, Yu Pei, Siyu Long, Yuxuan Song, Zhe Zhang, Wenhao Huang, Ziyao Cao, Shuyi Zhang, Hao Zhou, and Wei-Ying Ma. Steering protein family design through profile bayesian flow. *arXiv preprint arXiv:2502.07671*, 2025.

Changze Lv, Jiang Zhou, Siyu Long, Lihao Wang, Jiangtao Feng, Dongyu Xue, Yu Pei, Hao Wang, Zherui Zhang, Yuchen Cai, et al. Amix-1: A pathway to test-time scalable protein foundation model. *arXiv preprint arXiv:2507.08920*, 2025.

Rongqing Yuan, Jing Zhang, Andriy Kryshtafovych, R Dustin Schaeffer, Jian Zhou, Qian Cong, and Nick V Grishin. Casp16 protein monomer structure prediction assessment. *Proteins: Structure, Function, and Bioinformatics*, 94(1):86–105, 2026.

Kolja Stahl, Andrea Graziadei, Therese Dau, Oliver Brock, and Juri Rappsilber. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature biotechnology*, 41(12):1810–1819, 2023.

Alisia Fadini, Minhuan Li, Airlie J McCoy, Suresh Banjara, Hiroki Okumura, Eve Napier, Pietro Fontana, Amir R Khan, Luca Jovine, Thomas C Terwilliger, et al. Alphafold as a prior: experimental structure determination conditioned on a pretrained neural network. *Nature methods*, pages 1–11, 2026.

Feng Yu, Stephanie Prince, Andrew Tritt, Kanupriya Pande, Greg L Hura, Oliver Ruebel, and Susan E Tsutakawa. Experimental data driven ai framework for flexible protein conformational reconstruction. *bioRxiv*, 2026.

Advaith Maddipatla, Anar Rzayev, Marco Pegoraro, Martin Pacesa, Paul Schanda, Ailie Marx, Sanketh Vedula, and Alex M Bronstein. Inference-time optimization for experiment-grounded protein ensemble generation. *arXiv preprint arXiv:2602.24007*, 2026.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Timothy Truong Jr and Tristan Bepler. Poet: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36:77379–77415, 2023.

Tamas Lazar, Elizabeth Martínez-Pérez, Federica Quaglia, András Hatos, Lucía B Chemes, Javier A Iserte, Nicolás A Méndez, Nicolás A Garrone, Tadeo E Saldaño, Julia Marchetti, et al. Ped in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic acids research*, 49(D1):D404–D411, 2021.

Hsin-Hui Wu, Jindrich Symersky, and Min Lu. Structure and mechanism of a redesigned multidrug transporter from the major facilitator superfamily. *Scientific Reports*, 10(1):3949, 2020.

Kumar Nagarathinam, Yoshiko Nakada-Nakura, Christoph Parthier, Tohru Terada, Narinobu Juge, Frank Jaenecke, Kehong Liu, Yunhon Hotta, Takaaki Miyaji, Hiroshi Omote, et al. Outward open conformation of a major facilitator superfamily multidrug/h+ antiporter provides insights into switching mechanism. *Nature communications*, 9(1):4005, 2018.

Byung-Ha Oh, GF Ames, and Sung-Hou Kim. Structural basis for multiple ligand specificity of the periplasmic lysine-, arginine-, ornithine-binding protein. *Journal of Biological Chemistry*, 269 (42):26323–26330, 1994.

Heikki Takala, Alexander Björling, Oskar Berntsson, Heli Lehtivuori, Stephan Niebling, Maria Hoernke, Irina Kosheleva, Robert Henning, Andreas Menzel, Janne A Ihalainen, et al. Signal amplification and transduction in phytochrome photosensors. *Nature*, 509(7499):245–248, 2014.

Christian Köhler, Janet K Lighthouse, Tobias Werther, Olav M Andersen, Annette Diehl, Peter Schmieder, Jianguang Du, Bernadette C Holdener, and Hartmut Oschkinat. The structure of mesd45–184 brings light into the mechanism of ldlr family folding. *Structure*, 19(3):337–348, 2011.

Alex S Holehouse and Birthe B Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature reviews Molecular cell biology*, 25(3):187–211, 2024.

Mustapha Carab Ahmed, Ramon Crehuet, and Kresten Lindorff-Larsen. Computing, analyzing, and comparing the radius of gyration and hydrodynamic radius in conformational ensembles of intrinsically disordered proteins. In *Intrinsically disordered proteins: methods and protocols*, pages 429–445. Springer, 2020.

# A Mathematical Explanation

## A.1 Distribution shift of row addition

**Proposition 2** *Row addition operation shifts the empirical distribution of the MSA:*

$$p_{M'} = \frac{N}{N+\Delta N} p_M + \frac{\Delta N}{N+\Delta N} p_{\widetilde{M}} \tag{12}$$

**Proof. of Proposition 2** Let the empirical distribution induced by an MSA $M \in \Sigma^{N \times L}$ be defined as:

$$p_M(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(M_{i,:} = x),$$

where $x \in \Sigma^L$ denotes a sequence and $\mathbf{1}(\cdot)$ is the indicator function.

After applying the row-adding operation, we obtain:

$$M' = \begin{bmatrix} M \\ \widetilde{M} \end{bmatrix}, \quad \widetilde{M} \in \Sigma^{\Delta N \times L}.$$

Thus, the empirical distribution of $M'$ becomes:

$$p_{M'}(x) = \frac{1}{N+\Delta N} \left( \sum_{i=1}^{N} \mathbf{1}(M_{i,:} = x) + \sum_{j=1}^{\Delta N} \mathbf{1}(\widetilde{M}_{j,:} = x) \right).$$

Rearranging terms, we obtain:

$$p_{M'}(x) = \frac{N}{N+\Delta N} \cdot \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(M_{i,:} = x) + \frac{\Delta N}{N+\Delta N} \cdot \frac{1}{\Delta N} \sum_{j=1}^{\Delta N} \mathbf{1}(\widetilde{M}_{j,:} = x).$$

Therefore,

$$p_{M'}(x) = \frac{N}{N+\Delta N} p_M(x) + \frac{\Delta N}{N+\Delta N} p_{\widetilde{M}}(x).$$

This shows that the new distribution is a convex combination of the original MSA distribution and the generated sequence distribution, which completes the proof. □

## A.2 Coverage of Single-Action Space

**Explanation of Proposition 1** In this section, we provide a constructive justification for Proposition 1 that the proposed composed action space subsumes commonly used single-operation transformations.

We consider four representative classes of single operations: MSA shadowing (row selection), MSA clustering (row clustering), column masking, and amino acid (AA) mutation. We show that each of them can be covered or strictly generalized by our primitive actions.

**Row operations: shadowing and clustering.** Both MSA shadowing and MSA clustering operate along the sequence (row) dimension. Shadowing explicitly selects a subset of sequences, while clustering reduces redundancy by grouping similar sequences and keeping representatives.

Our row deletion operator $a_{\text{row-del}}$ directly performs subset selection:

$$M' = M[\mathcal{I}, :], \quad \mathcal{I} \subset \{1, \ldots, N\}$$

which is sufficient to simulate shadowing. Moreover, clustering can be viewed as selecting representative sequences from clusters, which can also be approximated by appropriate row selection. By offering a single operator $a_{\text{row-del}}$, language model is capable to conduct clustering, filtering, representative detecting and other options which could be implemented by coding in virtual environment.

**Column operation: masking.** MSA masking modifies residue identities at specific positions across sequences, which corresponds to operations along the column dimension. Our column masking operator $a_{\text{col-mask}}$ applies masking at the full-column level:

$$M'_{i,j} = \begin{cases} X, & j \in \mathcal{J} \\ M_{i,j}, & \text{otherwise} \end{cases}$$

This directly captures the effect of column-wise masking. Iterative applications further allow flexible masking patterns, hence covering standard MSA masking procedures.

**Amino acid mutation via generative modeling.** Single or multiple amino acid mutations modify residue identities at selected positions within sequences. Instead of explicitly defining mutation operators, we leverage the row adding operator $a_{\text{row-add}}$, which samples new sequences from a protein language model:

$$\widetilde{M} \sim p_{\text{PLM}}(\cdot \mid M)$$

Modern protein language models used for protein family design are trained to generate diverse yet plausible sequences conditioned on an input MSA. As a result, the generated sequences implicitly explore a mutation space that is strictly larger than predefined single-site or multi-site mutation rules.

Therefore, AA mutation operations can be viewed as a restricted subset of the sequence variations induced by $a_{\text{row-add}}$.

Note that we only claim that the designing space of protein language model covers the space of single residue variants, and in most times, explores large mutation space than multiple step direct mutation methods. There may exists mutations that can't generated by protein language model(e.g. completely different sequence). However, those sequence are not considered useful information to expand current protein family.

### A.3 Composition of actions

At each iterative searching step, we claim a composition of three MSA operations is conducted on current MSA state in Equation 5: $a = a_{\text{col-mask}} \circ a_{\text{row-add}} \circ a_{\text{row-del}}$. Here we wanted to clarify that any MSA transformation $T$ is conducted on composed action $a$. However, language model can explicitly choose $\mathcal{I} = \{1, \ldots, N\}$ or $\mathcal{J} = \emptyset$, resulting in no effect during $a_{row-del}$ or $a_{col-mask}$. For $a_{row-add}$, we explicitly restrict $\Delta N \leq 0.1N$ to avoid rapid distributional shift in single step.

## B More Related Work

### B.1 Recursive Language Model

Recursive Language Model(RLM) Zhang et al. [2025] is a general inference strategy where language models can decompose and recursively interact with their long input context as a variable. RLM releases more potential of language model by programmatically switching context library in arbitrary REPL workflows. As a variants of Test-Time-Scaling, RLM allows models to think at overall global level and make call-backs to smaller LLM, resulting in new schemes of chunking and adaptively recursion. In our study, RLM would serve to be an important backbone, ensembling model to handle long MSA context and tool calls.

### B.2 Tree of Thought

First proposed by [Yao et al., 2023], tree of thoughts is a searching paradigm through a combinatorial problem space. Unlike single sequential chain of thought(CoT), tree of though(ToT) allows language models to explore multiple reasoning paths while keeping a balance between reasoning efficiency and depth. In addition to theoretical backgrounds, the framework of ToT has been used to solve downstream applications like suzudo puzzles[Long, 2023], language QAs[Mo and Xin, 2024] [Silva et al., 2024] and creative writings[Yao et al., 2023]. In this study, we found that the "tree of thought" concept much aligned to the recursive nature of RLM.

### B.3 Protein In-Context Learning

Large Language Models have the capability to learn from a few highly related context examples, so does protein language model and structure prediction model. Both protein understanding designing benefits from in-context information, especially for sequence-level retrieved MSAs [Wu et al., 2025] and structure-level templates. Therefore, researchers has been prompting structure prediction models like AlphaFold and ESMFold with manually enhanced co-evolutionary contexts[Gong et al., 2025] [Lv et al., 2025]. This seems to be an alternative way than model parameter scaling in application scenario of specific protein.

### B.4 MSA-based method in conformation prediction

A growing line of work has explored MSA manipulation strategies to improve structure prediction and promote conformational diversity. Early approaches focus on filtering and pruning MSAs based on sequence similarity or confidence [Zhang et al., 2026, Yuan et al., 2026] scores to reduce redundancy. Subsequent methods, such as MSA subsampling [Del Alamo et al., 2021] and AF-cluster [Wayment-Steele et al., 2024], demonstrate that carefully selected subsets of sequences can retain predictive power while enabling the discovery of alternative conformations. Other approaches, including the AFSample series [Kalakoti and Wallner, 2025, 2026], introduce stochastic perturbations such as column masking to weaken dominant evolutionary signals and encourage structural diversity.

**Experimental-data-guided optimization of AlphaFold latent representations.** A complementary line of work incorporates non-MD experimental measurements into the latent or intermediate representations of protein structure-prediction models to steer inference toward experimentally supported conformations. AlphaLink [Stahl et al., 2023] incorporates photo-crosslinking mass-spectrometry restraints into OpenFold's pair representation, allowing the Evoformer to combine sparse experimental contacts with coevolutionary information from the MSA. [Fadini et al., 2026] propose ROCKET, which directly optimizes AlphaFold2/OpenFold MSA cluster profiles against X-ray crystallographic and cryo-EM likelihoods to explore experimentally supported structural variation. [Yu et al., 2026] develop AlphaSAXS, which introduces SAXS-derived pair-distance distributions through dedicated attention modules acting on both the MSA and pair representations of an AlphaFold-like architecture. [Maddipatla et al., 2026] further formulate experiment-grounded ensemble generation as inference-time optimization over AlphaFold3 latent representations, combining structural and force-field priors with experimental likelihoods and Boltzmann-weighted sampling. Collectively, these studies demonstrate that experimental observations can reshape the intermediate representations of pretrained structure-prediction models.

## C Detailed Implementation of MSA-Evolver

### C.1 Recursive Sub-Calls

To conduct multi-step reasoning and guidance of environmental feedback, MSA-Evolver's MSA searching process is implemented with Recursive Language model [Zhang et al., 2025]. As described in Appendix B, RLM brought convenience to MSA-operation language model: 1) Directly MSA editing in REPL environment: MSA is constructed as a local long variable. 2) Recursive sub-calls: the language model can complete code like this to spawn a sub-RLM, which recursively considers the current MSA as a starting point and keeps the searching process running until the max depth is reached.

```
result = rlm_query(prompt= "....")
```

where sub-rlm wil continues editing on local variable MSA matrix. When max-depth reached, *rlm_query* degenerated into simple *llm_query*.

### C.2 PLM Tools

As implement for action row-adding, protein language model is involved to generate virtual MSA based on current selected MSA. The task is considered as protein family design, which perform new protein sequence similar to existing sequences(MSA) as condition. Among protein language

**Algorithm 1** MSA-Evolver Conformation Search

**Require:** Initial MSA $M_0$ , Language model $p_\theta$, MSA state generator $G(\cdot)$, Candidate number $k$, Folding model $p_F$, Historical information $\mathcal{H}$, Anchor Set $\mathcal{B}$ Maximum searching depth $D$, Total folding budgets $n$, RMSD threshold $\delta$

**Ensure:** Full search trajectory $\mathcal{T} = [M_0, M_1, \ldots, M_n]$

Initialize $\mathcal{T} \leftarrow [s_0]$, $s_0 = M_0$, $y_0 \sim p_F(\cdot \mid M_0)$, $c \leftarrow 1$, $\mathcal{H} \leftarrow \{(\emptyset, 0)\}$, $\mathcal{B} \leftarrow \{y_0\}$
Initialize depth-prioritized queue $\mathcal{Q}$ (sorted by: 1. depth descending; 2. $V(s)$ descending)
Initialize state: push$(\mathcal{Q}, (s_0, \text{depth} = 0, V(s_0) = 0))$
**while** $c < n$ **and** $\mathcal{Q} \neq \emptyset$ **do**
  $(\hat{s}, d_{\hat{s}}, V_{\hat{s}}) \leftarrow \text{pop}(\mathcal{Q})$
  **if** $d_{\hat{s}} \geq D$ **then**
    **continue**
  **end if**
  $\mathcal{C} \leftarrow G(p_\theta, \hat{s}, k, \mathcal{H})$
  **for all** $s \in \mathcal{C}$ **do**
    $y_s \sim p_F(\cdot \mid s)$
    $V_s = \min_{y' \in \mathcal{B}} RMSD(y_s, y')$
    $d_s = d_{\hat{s}} + 1$
    push$(\mathcal{Q}, (s, d_s, V_s))$
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{s\}$
    $c \leftarrow c + 1$
    $\mathcal{H} \leftarrow \mathcal{H} \cup (s, V_s)$
    **if** $V_s \geq \delta$ **then**
      $\mathcal{B} \leftarrow \mathcal{B} \cup y_s$
    **end if**
    **if** $c \geq n$ **then**
      **Break**
    **end if**
  **end for**
**end while**
**return** Full search trajectory $\mathcal{T}$

models capable, we choose AMix-1.0 as a powerful example due to its systematic scaling law analysis assemble to LLM and strong in-silico and wet-lab verifications. In our study, AMix-1.0 is decorated as a sequence generating tool according to language model, which is :

```
v_MSA = plm_query(MSA= current_MSA , n_samples=n)
```

where $current_MSA \in \Sigma^{N \times L}$, $v_MSA \in \Sigma^{n \times L}$

# D Algorithm

MSA-Evolver's searching exploration algorithm is shown in Algorithm. 1. To balance the extent of searching depth and width, beam search is considered using a prioritized queue structure.

# E Datasets

## E.1 Custom dataset for conformation changing pairs

We used MSAs offered by [Kalakoti and Wallner, 2025], otherwise MSA are retrived from UniRef100 [Suzek et al., 2007] dataset) under ColabFold [Mirdita et al., 2022] pipeline. After collection of all protein conformation pairs from [Jing et al., 2023], [Lu et al., 2024] and [Kalakoti and Wallner, 2025], we remove a few protein targets due to lack of searched MSA sequences(less than 10 matches in UniRef100 [Suzek et al., 2007] dataset). For proteingym benchmark, all protein MSAs undergoes a similarity pre-filtering stages following [Truong Jr and Bepler, 2023]. MSA aligned sequence that have less than 10% similarity with protein target are removed to ensure MSA quality.

Table 5: Ablation study on restricted action space. Each action positively contributes to overall searched trajectory in MSA Space.

| Method | Action Space | | | Open-Close | | In-out | | Apo-Holo | | Fold-Switch | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Row-del. | Col-mask | Row-add. | Pref. | Alter. | Pref. | Alter. | Pref. | Alter. | Pref. | Alter. |
| MSA-Evolver | ✓ | | | **0.941** | 0.841 | 0.855 | 0.625 | 0.888 | 0.792 | 0.800 | 0.650 |
| | ✓ | ✓ | | 0.936 | 0.849 | 0.849 | 0.651 | **0.895** | 0.793 | 0.851 | **0.683** |
| | ✓ | ✓ | ✓ | 0.940 | **0.854** | **0.861** | **0.665** | 0.888 | **0.797** | **0.852** | **0.683** |

For all proteins and their PDB-IDs, please refer to Table 8, 9, 10, 11, 12, and 13 .

### E.2 PED data for IDPs

Following [Lu et al., 2024], we import IDPs from PED database [Lazar et al., 2021]. Among all 115 proteins, we filtered it to 103 targets after removal orphan sequence and those which fails to endure AF_cluster [Wayment-Steele et al., 2024] pipeline, that is PED00009, PED00166, PED00204, PED00205, PED00278, PED00294, PED00302, PED00303, PED00348, PED00393, PED00493, PED00500.

## F Experimental Details

### F.1 Baselines

Detailed information about metrics, reproduce links and applications are shown in Table 6.

Table 6: Detailed information and implement link for baseline methods

| Method | Metrics | Links | Application | Reference |
|---|---|---|---|---|
| ProtTeX | ROUGE-L, TM-score, RMSD | | Protein QA, Conformation Prediction, Function Annotation | [Ma et al., 2025] |
| ESM-Diff | TM-score, RMSD, JS div., MAE | | Conformation Prediction of apo-holo fold-switch protein pairs and IDP | [Lu et al., 2024] |
| AF_cluster | RMSD, plDDT, PCA | | Condormation Prediction of fold-switch proteins | [Wayment-Steele et al., 2024] |
| MSA-Subs. | TM-score, RMSF, RMSD | | Conformation Prediction of transporters and G-protein coupled receptors | [Del Alamo et al., 2021] |
| AFSample2 | TM-score, RMSF | | conformational change of open-closed state and transport proteins | [Kalakoti and Wallner, 2025] |
| AFSample3 | Pref.TM, Alter.TM | | General Conformation Prediction | [Kalakoti and Wallner, 2026] |

### F.2 Hyper-parameters

Each protein targets are expected to generate 10 ensembles for all experiments. For MSA-Evolver's search itself, max MSA depth starts with 15,360, which is 3 times of max MSA depth from MSA-Subs. [Del Alamo et al., 2021] to enable boarder space search. Minimal MSA depth is set to 16 to avoid degeneration to single sequence. All attempts that violates depth regulation are leaf-cut.

For protein language model, that is, AMix-1.0-650M [Lv et al., 2025], generate config as follows:

- **sample_t = 0.7** For a model of protein family design, $t \in [0, 1]$ controls of how similar generated sequence according to input MSA. Default set to 0.6, here we think 0.7 is enough to ensure virtual MSA quality in practice.
- **batch_size = 64** A normal batch size to form a stable tool.
- **sample_step = 500** Default sample step recommanded by [Gong et al., 2025].

The default searching depth of recursive language model is set to 5. In IDP experiemnts, we set depth=1 to encourage model to kept exploring on bigger MSAs.

For RMSD threshold $\delta$ when updating candidates set, we use 1.0 $\mathring{A}$ for BPTI experiments while default value of 5.0 $\mathring{A}$ for other experiments.

### F.3 Computational Resources

All experiment conducted on NVIDIA-A100 with cuda version 12.8, including folding based on Protenix [Team et al., 2025](an open-source reproduction of AlphaFold3) and vitual MSA generating based on AMix-1.0 [Lv et al., 2025]. For Language model, we use API calling for Claude-Sonnet-4.6.

### F.4 Evaluation metrics

**RMSD(Root-Mean-Square Deviation)** For a protein of $N$ residues, given predicted coordinates ($\{x_i\}_{i=1}^N$) and reference coordinates ($\{y_i\}_{i=1}^N$), RMSD is computed as:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|x_i - y_i|^2}.$$

**TM-score(Template Modeling Score)**

$$\text{TM-Score} = \max\left[\frac{1}{L_{\text{ref}}}\sum_{i=1}^{L_{\text{ali}}}\frac{1}{1+(d_i/d_0(L_{\text{ref}}))^2}\right],$$

where $d_i$ is the distance between aligned residues, $L_{\text{ref}}$ is the length of the reference protein, $L_{\text{ali}}$ is the number of aligned residues, and $d_0$ is a length-dependent normalization factor.

In this study, we calculate Best of n value of TM-score following baselines. This is an oracle-style evaluation intended to measure whether a method can sample a target conformational state within a fixed folding budget. It does not address the separate problem of selecting the biologically relevant alternative conformation without experimental references.

**Mean absolute error (MAE)** : For each feature, we calculate the ensemble average for both the reference and generated samples, then compute the MAE between them and average this value over the channels.

Table 7: Ablation study on search strategy. Historical information and LLM decision contributions to MSA-Evolver's effectness.

| Method | Open-Close | | In-out | | Apo-Holo | | Fold-Switch | |
|---|---|---|---|---|---|---|---|---|
| | Pref. | Alter. | Pref. | Alter. | Pref. | Alter. | Pref. | Alter. |
| DFS | 0.933 | **0.856** | 0.846 | 0.648 | 0.885 | 0.804 | 0.832 | 0.666 |
| BFS | 0.931 | 0.849 | 0.855 | **0.668** | 0.887 | 0.795 | 0.843 | 0.673 |
| **MSA-Evolver** | **0.940** | 0.854 | **0.861** | 0.665 | **0.888** | **0.806** | **0.856** | **0.684** |

**Pairwise distance** : Pairwise distances are calculated between atoms, skipping three atoms between pairs. Following [Lu et al., 2024], histograms with $N_{bin} = 50$ and pseudocount of $\epsilon = 10^{-6}$ to avoid zero frequencies for slight smoothing.

**Radius of gyration** : This measures the root mean square distance of atoms from the center of mass. It quantifies how far, on average, the atoms are distributed from the center of mass of protein structure.

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i \left\|\mathbf{r}_i - \mathbf{r}_{\text{cm}}\right\|^2}{\sum_{i=1}^{N} m_i}}$$

where $N$ is the number of atoms, $m_i$ is the mass of atom $i$, $\mathbf{r}_i$ is the position vector of atom $i$, and $\mathbf{r}_{\text{cm}}$ is the center of mass:

$$\mathbf{r}_{\text{cm}} = \frac{\sum_{i=1}^{N} m_i \mathbf{r}_i}{\sum_{i=1}^{N} m_i}.$$

**Contact map** : Contact maps are calculated as the probability distribution over a 2D grid of pairwise distances. After calculation of pair-wise distance, a threshold of 8.0 Åis applied to determine if two residues are in contact. As a result, the frequency of contacts across all conformations in the ensemble is normalized to create a probability distribution over the entire distance map. The contact probability between residues i and j is given by:

$$P_{contact}(i,j) = \frac{1}{N}\sum_{k=1}^{N}\mathbf{1}[d^{(k)}(i,j) \leq 8.0\mathring{A}]$$

where $d^{(k)}(i,j) \in \mathbb{R}$ is the pairwise distance between residues $i$ and $j$ in conformation $k$, and $N$ is the number of conformations in the ensemble.

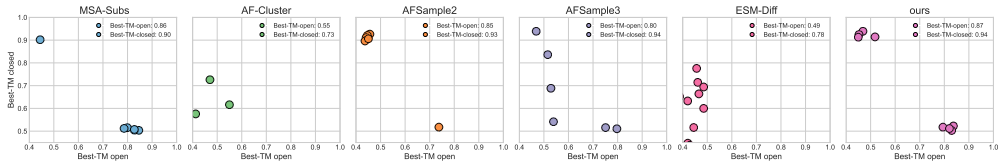



Figure 3: Case Study of protein SLC1A1 presented using TM-score axis. Among Folding samples, MSA-Evolver accurately found conformation hit points.

### F.5 Case Study of Conformation Switching Pairs

**MdfA (6VS1_A and 6GV1_A)** The Escherichia coli multidrug transporter MdfA operates via a classic "rocker-switch" mechanism, actively pumping drugs out of the cell by alternating between extreme inward-outward conformations. Structural studies have successfully captured these two functional extremes: the inward-facing state (PDB: 6VS1_A), where the binding cavity opens to the cytoplasm to capture incoming drugs [Wu et al., 2020], and the outward-facing state (PDB: 6GV1_A), which opens to the periplasm to expel them [Nagarathinam et al., 2018]. Comparing these distinct structural states allows researchers to dissect how protonation drives the large-scale reorientation of the substrate-binding pocket during the efflux cycle. Understanding these dynamic transitions provides a critical theoretical foundation for designing specific inhibitors to combat the growing threat of bacterial antimicrobial resistance.

**LAO (1LAH_E and 2LAO_A)** The Salmonella typhimurium LAO binding protein employs a "venus flytrap" mechanism to capture specific amino acids and deliver them to inner membrane ABC transporters. This precise delivery relies on a dramatic hinge-bending motion that continuously transitions the protein between distinct apo-holo conformational states. In the ligand-free apo state (PDB: 2LAO_A), the protein domains are widely splayed to expose the binding pocket, whereas substrate entry induces a tight structural collapse into the ligand-bound holo state (PDB: 1LAH_E) [Oh et al., 1994]. Analyzing these structural extremes not only elucidates the induced-fit dynamics of high-affinity substrate capture, but also explains how this closed conformation functions as a critical trigger for downstream signaling cascades.

**Bacteriophytochrome Photosensory Core Module (4O0P_A and 4O01_D)** The Deinococcus radiodurans bacteriophytochrome photosensory core module acts as a "light-controlled mechanical switch" that translates photon absorption into macroscopic cellular signals via dramatic open-closed conformational shifts. Structural analyses have captured this spatial signal amplification mechanism, starting from a compact, dark-adapted closed resting state (PDB: 4O0P_A) [Takala et al., 2014]. Upon red-light illumination, a minuscule atomic isomerization of the chromophore cascades into the refolding of an adjacent "tongue" region, culminating in the globally splayed, light-activated open state (PDB: 4O01_D) [Takala et al., 2014]. Deciphering how this localized photochemical event drives massive long-range structural rearrangements is vital for understanding photoreceptor signaling and designing precise near-infrared optogenetic tools.

**MESD (2RQK_A and 2RQM_A)** The Mus musculus mesoderm development protein (MESD) is a specialized endoplasmic reticulum chaperone that utilizes an auto-regulatory fold-switching mechanism to modulate its own activity. NMR studies reveal that in its inactive resting state (PDB: 2RQK_A), a flexible N-terminal $\alpha$-helix folds back to mask a sticky hydrophobic core, thereby preventing non-specific protein aggregation [Köhler et al., 2011]. To assist in the folding of low-density lipoprotein receptors, this helix undergoes a massive spatial displacement, transforming into an unmasked, active conformation (PDB: 2RQM_A) that securely binds incompletely folded polypeptides [Köhler et al., 2011]. Analyzing this dramatic secondary structure rearrangement highlights how MESD dynamically exposes its binding interface to maintain protein homeostasis and regulate essential developmental signaling pathways.

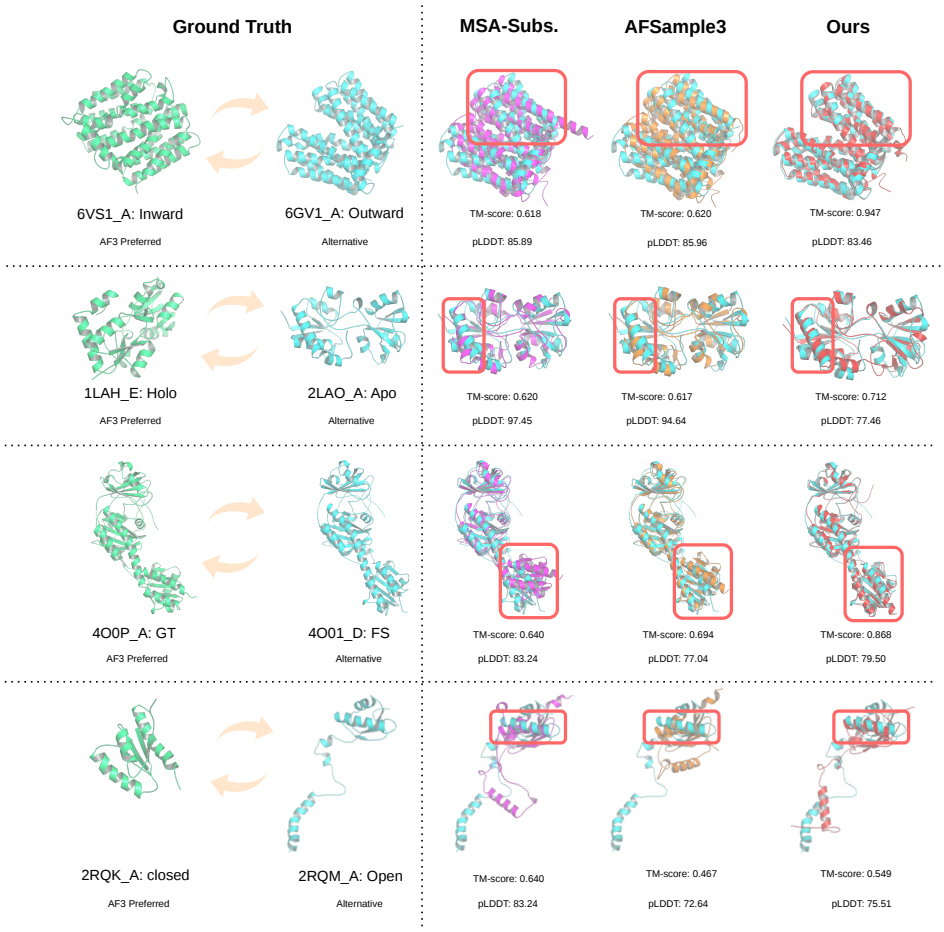



Figure 4: Case Study of conformation changing pairs.

### F.6 IDP

Intrinsically Disordered Proteins (IDPs) are proteins that lack a stable tertiary structure under physiological conditions [Oldfield and Dunker, 2014, Trivedi and Nagarajaram, 2022]. Unlike globular proteins, IDPs do not fold into a unique three-dimensional conformation. Instead, they exist as ensembles of rapidly inter-converting conformations, which enable them to participate in diverse biological processes such as signaling, regulation, and molecular recognition [Holehouse and Kragelund, 2024].

Predicting the conformational ensembles of IDPs is crucial for understanding their biological function and interactions. Since IDPs do not adopt a fixed structure, traditional structure prediction approaches that aim for a single static conformation are insufficient. Instead, computational methods focus on generating ensembles of structures that can recapitulate experimentally observable properties such as pariwise distance and radius of gyration [Ahmed et al., 2020].

### F.7 BPTI

The conformational dynamics of bovine pancreatic trypsin inhibitor (BPTI) have been extensively characterized through long-timescale molecular dynamics simulations, where five kinetically distinct conformational clusters have been identified [Huber et al., 1974, Brooks and Karplus, 1983]. Owing to this well-established structural landscape, BPTI serves as a representative benchmark for evaluating the capability of models to capture multiple conformational states. Among these clusters, Cluster 3

Table 8: Protein Conformational Change Dataset: Open-Closed

| Dataset Type | Conf.1(Open) | Conf.2(Closed) | Length | TM-score | RMSD |
|---|---|---|---|---|---|
| | 3FTO_A | 3DRF_A | 576 | 0.507 | 4.66 |
| | 4BP8_A | 4BP9_A | 714 | 0.523 | 4.31 |
| | 6YED_A | 6YE0_A | 343 | 0.565 | 3.79 |
| | 2XE6_A | 2WZD_A | 416 | 0.598 | 3.62 |
| | 3TEE_A | 3VJP_A | 197 | 0.612 | 2.48 |
| | 3L6G_A | 3L6H_A | 253 | 0.617 | 3.34 |
| | 7QGA_A | 4H2I_A | 522 | 0.621 | 4.27 |
| | 7CY2_A | 7CYR_A | 438 | 0.633 | 2.80 |
| | 3ZSF_A | 2YLN_A | 258 | 0.636 | 3.45 |
| | 3IUJ_A | 3IUQ_A | 689 | 0.648 | 3.33 |
| | 6HNJ_A | 6HNI_A | 316 | 0.656 | 3.01 |
| | 2RQM_A | 2RQK_A | 139 | 0.664 | 1.60 |
| | 2KTV_A | 2KTU_A | 161 | 0.708 | 3.12 |
| | 6MKA_A | 6MKJ_A | 639 | 0.720 | 3.46 |
| | 2WFA_A | 2WF5_A | 220 | 0.748 | 2.92 |
| | 3O6W_A | 3O8M_A | 484 | 0.749 | 3.14 |
| | 1SI1_A | 1SI0_A | 319 | 0.749 | 2.40 |
| | 6HAC_A | 6HAE_A | 341 | 0.754 | 2.88 |
| | 6NC7_A | 6NC6_A | 474 | 0.755 | 2.79 |
| | 5HO2_A | 5HO0_A | 847 | 0.766 | 3.45 |
| | 7C63_A | 7C66_A | 415 | 0.781 | 2.04 |
| | 6K8B_A | 6K85_B | 290 | 0.824 | 2.18 |
| Open_Closed | 2OLO_A | 2OLN_A | 388 | 0.837 | 1.76 |
| | 7CL7_A | 7CL8_A | 399 | 0.851 | 1.85 |
| | 2ZPD_A | 2ZPC_A | 181 | 0.855 | 1.91 |
| | 4UTO_A | 4UTP_A | 308 | 0.857 | 1.61 |
| | 4IRP_A | 4IRQ_A | 246 | 0.862 | 1.29 |
| | 2IWV_A | 2IWW_A | 279 | 0.884 | 2.18 |
| | 2WIO_A | 2JJN_A | 396 | 0.889 | 1.94 |
| | 3QUQ_A | 3QU2_A | 243 | 0.895 | 1.27 |
| | 5EQD_B | 5EQD_A | 399 | 0.898 | 1.68 |
| | 5K9V_A | 5K9W_A | 300 | 0.900 | 0.95 |
| | 6JMX_A | 6JMY_A | 253 | 0.908 | 1.67 |
| | 4HJP_A | 4HJQ_A | 285 | 0.912 | 1.60 |
| | 3NS8_A | 2N2K_A | 75 | 0.922 | 1.26 |
| | 4KLX_A | 6NNI_A | 158 | 0.923 | 1.08 |
| | 2FY7_A | 2FYC_B | 272 | 0.925 | 1.29 |
| | 6XCS_A | 6XCQ_A | 417 | 0.9380 | 1.15 |
| | 7P0N_A | 7NGE_A | 388 | 0.946 | 0.33 |
| | 3NC3_A | 3NC3_B | 404 | 0.949 | 1.16 |
| | 5AFA_A | 6TYR_A | 438 | 0.970 | 0.99 |
| | 3S9B_A | 3S9A_A | 235 | 0.986 | 0.44 |
| | 6QSM_A | 6QSL_A | 216 | 0.987 | 0.40 |
| | 4PSP_A | 4NI3_A | 584 | 0.990 | 0.47 |
| | 3M1Q_A | 3M1W_A | 259 | 0.996 | 0.29 |

corresponds to a remote folding mode that is particularly challenging to reproduce and is rarely captured by conventional folding approaches.

## G Limitations

Like other MSA-based methods, MSA-Evolver depends on the availability and quality of homologous sequences. It may be less effective for orphan proteins or proteins with shallow MSAs. However,

Table 9: Protein Conformational Change Dataset: Inward-Outward

| Dataset Type | Conf.1(Inward) | Conf.2(Outward) | Length | TM-score | RMSD |
|---|---|---|---|---|---|
| Inward_Outward | 6E9N_B | 6E9O_B | 409 | 0.773 | 4.18 |
| | 6GCI_A | 4C9J_B | 292 | 0.561 | 4.82 |
| | 6A6N_A | 7DQV_A | 587 | 0.788 | 4.30 |
| | 7MJS_X | 7N98_A | 458 | 0.756 | 4.54 |
| | 6FHZ_A | 3VVS_A | 440 | 0.736 | 4.73 |
| | 6S3Q_A | 6X3E_A | 430 | 0.621 | 4.92 |
| | 6XMS_A | 6XMU_A | 1091 | 0.759 | 6.21 |
| | 3QF4_A | 6QV1_C | 572 | 0.743 | 5.16 |
| | 4Q4A_B | 6QV1_B | 584 | 0.730 | 5.42 |
| | 5MKK_B | 6RAJ_B | 577 | 0.800 | 4.45 |
| | 6VS1_A | 6GV1_A | 387 | 0.775 | 4.20 |
| | 7L17_A | 4M64_A | 453 | 0.747 | 4.40 |
| | 6NC7_A | 6NC9_A | 469 | 0.675 | 5.60 |
| | 6BVG_A | 5IWS_A | 447 | 0.691 | 4.84 |
| | 6BTX_A | 5AYM_A | 377 | 0.670 | 5.31 |
| | 5C78_A | 6HRC_D | 564 | 0.669 | 4.99 |

we also recommend to manually have virtual MSA generated and use it as a virtual start-point for MSA-Evolver's MSA editing process, which would partly alleviate lack of evolutionary information.

# H Additional Methodological Details and Evaluation

## H.1 Initial MSAs and Structure Generation

MSA-Evolver starts from the complete searched MSA available for each target, without preprocessing it using another MSA-editing method. For the inward–outward transporter dataset (TP16) and the open–closed dataset (OC23), we use publicly released MSAs from the AFSample2 and AFSample3 studies. For the Apo–Holo and Fold–Switch datasets, initial MSAs are generated using the standard Protenix-Server[Team et al., 2025] MSA-search pipeline.

Starting from this initial alignment, MSA-Evolver explores a tree of MSA states by selecting an existing node and applying further editing operations. Every edited MSA generated within the folding budget is submitted to Protenix, the structure prediction module used in our experiments, and retained in the search trajectory. Its structural outcome is therefore observed through folding before it can inform subsequent exploration.

The anchor-structure set serves a separate purpose: maintaining a memory of structurally diverse predictions. Admission to this set depends on an RMSD threshold relative to existing anchors. This criterion governs anchor-set updates; it does not determine whether an edited MSA is folded or retained in the search trajectory.

## H.2 Search-Time Information and Reference-Based Evaluation

MSA-Evolver requires neither advance knowledge of whether a target exhibits multiple conformations nor a specified number of conformational states. Experimentally determined reference conformations are not supplied to the search. Instead, MSA editing is guided by the predicted structures and the resulting search history. The same procedure can therefore be applied to proteins without previously characterized alternative conformations.

Proteins with experimentally characterized conformations provide reference structures for evaluating conformational coverage. Given all predictions $\mathcal{Y}_i$ generated for target $i$ within the folding budget, the best-matching TM-score for reference conformation $r_{i,c}$ is

$$T_{i,c} = \max_{y \in \mathcal{Y}_i} \mathrm{TM}(y, r_{i,c}). \tag{13}$$

Preferred and alternative reference states are evaluated separately, following the same protocol for all methods. These scores assess whether a method samples each reference state within its budget.

Table 10: Protein Conformational Change Dataset: Apo-Holo Part. 1

| Dataset Type | Conf.1(Apo) | Conf.2(Holo) | Length | TM-score | RMSD |
|---|---|---|---|---|---|
| Apo_Holo | 1AEL_A | 1URE_A | 131 | 0.80656 | 2.23 |
| | 1BV2_A | 1UVB_A | 91 | 0.82641 | 1.69 |
| | 1C54_A | 1RGH_B | 96 | 0.89857 | 1.28 |
| | 1EAL_A | 1EIO_A | 127 | 0.83689 | 2.10 |
| | 1F3Y_A | 1JKN_A | 165 | 0.82528 | 2.60 |
| | 1FMF_A | 1ID8_A | 137 | 0.70031 | 3.11 |
| | 1GH1_A | 1CZ2_A | 90 | 0.78371 | 1.98 |
| | 1GQN_A | 1QFE_B | 252 | 0.98137 | 1.26 |
| | 1GQZ_A | 2GKE_A | 274 | 0.87630 | 2.45 |
| | 1HSI_B | 1HSH_D | 99 | 0.94221 | 0.99 |
| | 1I56_A | 1EL1_A | 130 | 0.86836 | 1.95 |
| | 1IGP_A | 2AU6_A | 175 | 0.97450 | 0.76 |
| | 1JFJ_A | 1JFK_A | 134 | 0.53534 | 3.66 |
| | 1K2H_A | 1ZFS_B | 93 | 0.52066 | 3.58 |
| | 1LIP_A | 1JTB_A | 91 | 0.67509 | 2.88 |
| | 1MUT_A | 1PUN_A | 129 | 0.65543 | 3.24 |
| | 1MX7_A | 1MX8_A | 134 | 0.91326 | 1.41 |
| | 1NTR_A | 1KRX_A | 124 | 0.71050 | 2.96 |
| | 1O1U_A | 1O1V_A | 127 | 0.88961 | 2.10 |
| | 1PDB_A | 1YHO_A | 186 | 0.93511 | 1.43 |
| | 1TFU_A | 3UC5_A | 157 | 0.96727 | 1.06 |
| | 1TJD_A | 1EEJ_B | 216 | 0.88661 | 2.08 |
| | 1XSA_A | 1XSC_A | 153 | 0.84167 | 2.50 |
| | 1Z15_A | 1Z17_A | 344 | 0.69137 | 4.07 |
| | 2CJO_A | 1ROE_A | 97 | 0.60864 | 3.31 |
| | 2D9E_A | 2RS9_B | 121 | 0.88665 | 1.38 |
| | 2F63_A | 1EQM_A | 158 | 0.82701 | 2.28 |
| | 2FHM_A | 2HLT_A | 91 | 0.89224 | 1.49 |
| | 2IN2_A | 2B0F_A | 182 | 0.90915 | 1.79 |
| | 2JU3_A | 2JU8_A | 127 | 0.75585 | 2.73 |
| | 2JWW_A | 1RTP_1 | 109 | 0.75840 | 2.46 |
| | 2K43_A | 2K8R_A | 133 | 0.74312 | 2.82 |
| | 2L50_A | 2L51_B | 102 | 0.77238 | 2.26 |
| | 2L68_A | 2LKK_A | 126 | 0.95856 | 0.88 |
| | 2LAO_A | 1LAH_E | 238 | 0.71294 | 3.80 |
| | 2LHS_A | 2BEM_A | 170 | 0.92676 | 1.42 |
| | 2NLN_A | 1RRO_A | 108 | 0.65718 | 3.33 |
| | 2P3M_A | 2VBT_A | 136 | 0.87068 | 2.03 |
| | 2UZ5_A | 2VCD_A | 137 | 0.78556 | 2.62 |
| | 4AKE_B | 2ECK_B | 214 | 0.68870 | 3.76 |
| | 1IJA_A | 2KID_A | 148 | 0.80326 | 2.60 |
| | 1JM4_B | 1WUM_A | 118 | 0.81256 | 2.35 |
| | 1MO7_A | 1MO8_A | 213 | 0.83160 | 2.87 |
| | 1SKT_A | 1TNQ_A | 90 | 0.57484 | 3.12 |
| | 1SYM_B | 1XYD_B | 92 | 0.58116 | 3.41 |
| | 1W4U_A | 1UR6_A | 147 | 0.76067 | 2.86 |
| | 2KXL_A | 2K0G_A | 142 | 0.72183 | 3.08 |
| | 2LKC_A | 2LKD_A | 178 | 0.80667 | 2.81 |
| | 1LMZ_A | 1P7M_A | 187 | 0.86741 | 2.34 |
| | 2CG7_A | 2CG6_A | 90 | 0.58395 | 2.31 |

Table 11: Protein Conformational Change Dataset: Apo-Holo Part. 2

| Dataset Type | Conf.1(Apo) | Conf.2(Holo) | Length | TM-score | RMSD |
|---|---|---|---|---|---|
| Apo_Holo | 1URP_D | 2DRI_A | 271 | 0.76218 | 3.36 |
| | 1ORM_A | 1QJ8_A | 148 | 0.56591 | 4.55 |
| | 1PFL_A | 1FIL_A | 139 | 0.89035 | 1.65 |
| | 1FSF_A | 1FQO_B | 266 | 0.96975 | 1.11 |
| | 1AKZ_A | 1SSP_E | 223 | 0.94441 | 1.50 |
| | 2AI6_A | 2OZW_A | 125 | 0.93468 | 1.45 |
| | 1VR6_A | 1RZM_A | 338 | 0.79766 | 2.05 |
| | 1Y3Q_A | 1Y3N_A | 490 | 0.77853 | 4.22 |
| | 1GUD_A | 1RPJ_A | 288 | 0.76355 | 3.65 |
| | 1EX6_B | 1EX7_A | 186 | 0.78325 | 3.15 |
| | 1WD7_B | 1WCW_A | 254 | 0.73038 | 3.66 |
| | 1W0J_E | 1E1R_F | 466 | 0.82826 | 3.59 |
| | 1RF5_A | 1RF4_A | 427 | 0.84492 | 2.99 |
| | 1S2O_A | 1TJ5_A | 244 | 0.81531 | 3.28 |
| | 1K5H_A | 1Q0Q_A | 398 | 0.86540 | 3.10 |
| | 1ZOL_A | 1O03_A | 221 | 0.83766 | 2.72 |
| | 1VIY_C | 1VHL_A | 208 | 0.89732 | 2.15 |
| | 1ZA1_A | 1Q95_A | 310 | 0.94372 | 1.63 |
| | 1JEJ_A | 1JG6_A | 351 | 0.92519 | 2.10 |
| | 1HOO_B | 1CG0_A | 431 | 0.94270 | 2.21 |
| | 1E5L_A | 1E5Q_A | 449 | 0.91684 | 2.44 |
| | 1HW1_B | 1H9G_A | 223 | 0.91871 | 1.67 |
| | 1L0W_B | 1G51_A | 580 | 0.96544 | 1.77 |
| | 1OTJ_D | 1GY9_A | 279 | 0.94633 | 1.74 |
| | 1EVK_A | 1EVL_A | 401 | 0.95845 | 1.64 |
| | 1G6W_D | 1K0B_C | 255 | 0.94716 | 1.36 |
| | 1NJG_B | 1NJF_A | 239 | 0.95377 | 1.33 |
| | 1RKA_A | 1GQT_A | 305 | 0.97034 | 1.22 |
| | 1K6W_A | 1K70_A | 423 | 0.98376 | 1.08 |
| | 1YL5_B | 1YL7_A | 245 | 0.95051 | 1.23 |
| | 6OY9_B | 1A0R_B | 339 | 0.94683 | 1.75 |
| | 1A6D_A | 1A6E_A | 503 | 0.99177 | 0.83 |
| | 2RCS_H | 1AJ7_H | 216 | 0.57441 | 3.36 |
| | 1AW2_A | 1AW1_B | 254 | 0.97952 | 1.02 |
| | 1RKM_A | 1B6H_A | 517 | 0.88495 | 3.03 |
| | 7AZP_A | 4PJ1_A | 524 | 0.63845 | 4.85 |
| | 4LP5_A | 4P2Y_A | 301 | 0.75727 | 3.52 |
| | 1CFC_A | 5DOW_A | 146 | 0.41114 | 4.79 |
| | 2KQ2_A | 2KW4_A | 147 | 0.33954 | 5.43 |
| | 1DMO_A | 3CLN_A | 148 | 0.31685 | 4.50 |

Table 12: Protein Conformational Change Dataset: Fold-Switch Part. 1

| Dataset Type | Conf.1 | Conf.2 | Length | TM-score | RMSD |
|---|---|---|---|---|---|
| Fold_Switch | 1K0N_A | 1RK4_B | 241 | 0.73532 | 2.58 |
| | 1MNM_C | 1MNM_D | 87 | 0.75257 | 2.07 |
| | 1NQJ_B | 1NQD_A | 119 | 0.86162 | 0.91 |
| | 1NOC_A | 1QOM_B | 388 | 0.87930 | 2.64 |
| | 1QS8_B | 1MIQ_B | 329 | 0.87043 | 2.64 |
| | 1REP_C | 2Z9O_B | 251 | 0.59390 | 4.49 |
| | 1UXM_K | 2NAM_A | 153 | 0.18512 | 5.02 |
| | 1X0G_A | 1X0G_D | 112 | 0.55323 | 2.36 |
| | 1XJU_B | 1XJT_A | 163 | 0.76203 | 2.32 |
| | 1XNT_A | 3LQC_A | 183 | 0.72217 | 2.90 |
| | 2HDM_A | 2N54_B | 92 | 0.30988 | 3.57 |
| | 2OUG_C | 6C6S_D | 162 | 0.59234 | 2.88 |
| | 3HDF_A | 3HDE_A | 140 | 0.77421 | 2.92 |
| | 3IFA_A | 5ET5_A | 338 | 0.83972 | 1.27 |
| | 3J7W_B | 3J7V_G | 345 | 0.65009 | 3.80 |
| | 3JV6_A | 1ZK9_A | 101 | 0.53404 | 2.32 |
| | 3KUY_A | 5C3I_F | 135 | 0.73420 | 1.80 |
| | 3M1B_F | 3LOW_A | 99 | 0.54691 | 1.70 |
| | 3O44_A | 1XEZ_A | 593 | 0.60135 | 3.55 |
| | 3QY2_A | 1QB3_A | 117 | 0.78319 | 0.75 |
| | 3TP2_A | 5LJ3_M | 229 | 0.81802 | 2.37 |
| | 3UYI_A | 3V0T_A | 337 | 0.85166 | 1.48 |
| | 3ZWG_N | 4TSY_D | 179 | 0.84118 | 0.83 |
| | 4AAN_A | 4AAL_A | 341 | 0.87190 | 2.58 |
| | 4GQC_C | 4GQC_B | 164 | 0.91414 | 1.35 |
| | 4J3O_F | 2JMR_A | 154 | 0.80276 | 1.96 |
| | 5HK5_H | 4JPH_B | 147 | 0.68506 | 2.45 |
| | 4NC9_C | 4N9W_A | 390 | 0.82927 | 2.77 |
| | 4O0P_A | 4O01_D | 523 | 0.84290 | 3.50 |
| | 4PHQ_A | 2WCD_X | 298 | 0.50032 | 4.64 |
| | 4PYI_A | 4PYJ_A | 221 | 0.88175 | 1.24 |
| | 4QHF_A | 4QHH_A | 110 | 0.64018 | 2.15 |
| | 4RWN_A | 4RWQ_B | 357 | 0.88668 | 2.56 |
| | 4UV2_D | 4Q79_F | 262 | 0.63563 | 2.59 |
| | 4Y0M_J | 4XWS_D | 227 | 0.88000 | 1.34 |
| | 4ZRB_C | 4ZRB_H | 137 | 0.92908 | 0.91 |
| | 5LY6_B | 5AOE_B | 471 | 0.47849 | 4.27 |

They do not require selecting two MSAs during search, assume that only two states exist, or solve the separate problem of identifying biologically relevant alternatives without experimental references.

### H.3 Confidence Analysis on Intrinsically Disordered Proteins

We examine whether the IDP results can be attributed to generating unusually low-confidence structures. We compare the average pLDDT of structures generated by MSA-Evolver, AFSample3, and MSA-Subsampling. We additionally pool generated samples across these methods and calculate Pearson and Spearman correlations between pLDDT and the downstream IDP error metrics: pairwise-distance (PWD) MAE, radius-of-gyration ($R_g$) MAE, and contact-map MAE.

The three methods have similar mean pLDDT values. All reported correlations are negative, with the strongest association for PWD MAE and the weakest for contact-map MAE. Thus, lower confidence is not associated with better IDP metrics in this analysis. These observations argue against an explanation based solely on producing low-confidence structures, although they do not establish a causal relationship between confidence and ensemble accuracy.

Table 13: Protein Conformational Change Dataset: Fold-Switch Part. 2

| Dataset Type | Conf.1 | Conf.2 | Length | TM-score | RMSD |
|---|---|---|---|---|---|
| Fold_Switch | 5C1V_A | 5C1V_B | 346 | 0.92330 | 1.04 |
| | 5FLU_E | 2UY7_D | 163 | 0.69028 | 3.41 |
| | 5HMG_A | 3ZTJ_E | 328 | 0.96356 | 0.54 |
| | 5I2M_A | 5I2S_A | 422 | 0.37327 | 4.47 |
| | 5JZH_A | 5JZT_G | 424 | 0.57162 | 3.44 |
| | 5KEQ_F | 1DZL_A | 483 | 0.89529 | 2.04 |
| | 5L35_D | 5L35_G | 422 | 0.89498 | 2.88 |
| | 3R9J_C | 2KXO_A | 77 | 0.41989 | 2.69 |
| | 1CEE_B | 2K42_A | 59 | 0.28474 | 3.72 |
| | 2H44_A | 1RKP_A | 326 | 0.88525 | 1.39 |
| | 2GRM_B | 2AXZ_A | 317 | 0.90578 | 1.21 |
| | 2C1V_B | 2C1U_C | 338 | 0.85690 | 2.49 |
| | 3KDS_G | 2CE7_C | 465 | 0.54137 | 3.17 |
| | 2GED_B | 1NRJ_B | 193 | 0.60810 | 2.18 |
| | 2LEL_A | 2K0Q_A | 74 | 0.54195 | 2.49 |
| | 2LV1_A | 6LNI_A | 147 | 0.13674 | 3.41 |
| | 2KKW_A | 2N0A_D | 140 | 0.18927 | 5.31 |
| | 1IYT_A | 2NAO_F | 42 | 0.22988 | 4.45 |
| | 2MWF_A | 2NNT_A | 32 | 0.20292 | 3.33 |
| | 1JFK_A | 2NXQ_B | 134 | 0.26832 | 4.16 |
| | 2P3V_D | 2P3V_A | 256 | 0.95562 | 1.18 |
| | 5JYT_A | 2QKE_E | 106 | 0.44649 | 3.53 |
| | 3G0H_A | 3EWS_B | 424 | 0.62219 | 5.10 |
| | 2VFX_L | 3GMH_L | 206 | 0.64991 | 2.40 |
| | 3J9C_A | 4H2A_A | 562 | 0.73689 | 2.99 |
| | 3NJQ_A | 2PBK_B | 193 | 0.84874 | 2.11 |
| | 3T1P_A | 1KCT_A | 371 | 0.81432 | 2.12 |
| | 4OW6_B | 4AE0_A | 535 | 0.64985 | 2.56 |
| | 4B3O_B | 3MEE_A | 441 | 0.45449 | 4.94 |
| | 4DXT_A | 4DXR_A | 198 | 0.91972 | 1.38 |
| | 4G0D_Z | 4FU4_C | 26 | 0.30892 | 1.98 |
| | 2LEP_A | 4HDD_A | 69 | 0.39627 | 1.46 |
| | 4RMB_B | 4RMB_A | 166 | 0.87174 | 1.12 |
| | 4YDQ_B | 4TWA_A | 497 | 0.94136 | 2.07 |
| | 5BMY_A | 5B3Z_A | 389 | 0.94130 | 2.05 |
| | 5EC5_P | 3ZXG_B | 298 | 0.57958 | 3.25 |
| | 5F5R_B | 5F3K_A | 238 | 0.72250 | 2.23 |
| | 5FHC_J | 1EBO_E | 98 | 0.24518 | 5.03 |
| | 7AHL_E | 4YHD_G | 293 | 0.79245 | 1.70 |
| | 6Z4U_A | 7KDT_B | 97 | 0.17355 | 3.72 |

Table 14: Mean prediction confidence on the IDP benchmark.

| Method | Mean pLDDT |
|---|---|
| MSA-Evolver | 78.21 |
| AFSample3 | 77.31 |
| MSA-Subsampling | 78.61 |

Table 15: Association between pLDDT and IDP evaluation errors across pooled samples. Lower MAE indicates better agreement with the reference ensemble.

| Metric | Pearson $r$ | Spearman $\rho$ |
|---|---|---|
| PWD MAE | $-0.550$ | $-0.643$ |
| $R_g$ MAE | $-0.431$ | $-0.462$ |
| Contact MAE | $-0.140$ | $-0.162$ |

### H.4 Performance on Additional Classification Task

**Setting.** We construct a binary classification benchmark combining TP16 as the positive set with the larger CAMEO-test used in [Gong et al., 2025] as the negative set. The task evaluates identification of known multi-conformational proteins against this background.

**Generation and scoring.** We compare MSA-Evolver, MSA-Subsampling[Del Alamo et al., 2021], and AFSample3[Kalakoti and Wallner, 2026]. Each method starts from the full MSA and generates up to 20 candidate structures per target. For TP16, the preferred reference $r_i$ is the experimentally characterized state corresponding to the target's starting conformation. For CAMEO-test, $r_i$ is the available experimental structure. These references are used for evaluation.

To reduce artifacts from severely distorted predictions, we retain a CAMEO-test candidate only when its TM-score to $r_i$ is at least $0.7$. Denoting the resulting candidate set by $\widetilde{\mathcal{Y}}_i$, we exclude targets with an empty set and calculate

$$s_i = \min_{y \in \widetilde{\mathcal{Y}}_i} \mathrm{TM}(y, r_i). \tag{14}$$

This target-level score corresponds to `min_tm_to_preferred_structure`. Lower values indicate that the retained ensemble contains a prediction farther from the preferred reference. Filtering is performed only during evaluation and leaves the generated predictions unchanged.

**Metrics and results.** We report AUPRC, computed by treating TP16 as the positive class and ranking targets by increasing $s_i$ (equivalently, using $-s_i$ as the positive-class score). We also report the absolute difference between the mean scores of retained CAMEO-test and TP16 targets, $\Delta_{\mathrm{mean}} = |\overline{s}_{\mathrm{CAMEO}} - \overline{s}_{\mathrm{TP16}}|$. Larger values indicate better ranking performance and greater mean separation, respectively.

Table 16: Classification results under the stated CAMEO-test filtering protocol, using up to 20 candidate structures per target. Best values are bolded.

| Method | AUPRC ↑ | $\Delta_{\mathrm{mean}}$ ↑ |
|---|---|---|
| MSA-Evolver | **0.740** | **0.193** |
| MSA-Subsampling | 0.525 | 0.031 |
| AFSample3 | 0.669 | 0.064 |

MSA-Evolver achieves the highest AUPRC and the largest mean separation under this protocol. These results support the utility of its generated structural ensembles for prioritizing known multi-conformational targets in this mixed benchmark.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: see Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: see Section G

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Refer to Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to Section. 2, 3 and Appendix. F

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Codes will be provided soon.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: see Appendix E and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Results are already mean value of various protein targets and structure ensembles.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: see Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is conducted on public datasets without model training.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 6.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such release.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: all public accessible assets have been cited in paper, that is, [Kalakoti and Wallner, 2025], [Jing et al., 2023] and [Lu et al., 2024].

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Datasets included in this study listed in Appendix E and detailed information about reproducing baselines are provided in Appendix 6.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: See Section. 2.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.