Sampling protein structural token space enables accurate prediction of multiple conformations

Zhiyuan Wang[1,2], Yue Yu[1,2], Chungong Yu[1,2,3*], Dongbo Bu[1,2,3*]

[1] SKLP, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China .

[2] University of Chinese Academy of Sciences, Beijing, 100049, China .

[3] Central China Artificial Intelligence Research Institute, Henan Academy of Sciences, Zhengzhou, 450046, China .

*Corresponding author(s). E-mail(s): chungongyu@gmail.com; dbu@ict.ac.cn;

Abstract

Protein function is fundamentally mediated by ensembles of distinct metastable states. However, existing methods typically exhibit a bias toward predicting a single dominant state, failing to capture alternative conformations or provide robust metrics for identifying high-quality multi-state conformations. Here, we present MultiStateFold (MSFold), a framework that integrates Parallel Tempering into the discrete structure token space of the ESM3 protein language model. By conceptualizing the model's latent space as an implicit energy landscape, MSFold enables global exploration and barrier crossing, thereby overcoming the local sampling limitations inherent in base generative models. Across a benchmark of 313 multi-conformation pairs, MSFold sets a new performance standard: it achieves the highest success rate in modeling native states and substantially outperforms leading methods, including AlphaFold 3 and MSA cluster, on challenging alternative conformations, while maintaining competitive accuracy for primary structures. Furthermore, we propose Sequence Log-Likelihood (SLL), a novel confidence metric derived from sequence-structure consistency. Our results demonstrate that SLL offers a modest improvement over standard metrics such as pTM and pLDDT. This work establishes a new paradigm for conformational sampling, bridging classical statistical physics with protein language models.

Keywords: Protein Multiple Conformations, Parallel Tempering, Protein Language Model, ESM3, Conformational Sampling

1 Introduction

Proteins are inherently dynamic macromolecules that function via ensembles of distinct metastable states [1]. While traditional structural biology often relies on static snapshots, proteins in solution populate a continuous spectrum of conformations. Enzymatic catalysis exemplifies this dynamism, requiring precise interconversion between states—such as open and closed forms—to coordinate substrate binding, catalysis, and product release. Beyond enzymes, conformational switching underpins diverse biological mechanisms, including molecular transport and signal transduction. These functional excursions span vast spatiotemporal scales, ranging from local side-chain fluctuations to global domain rearrangements. Consequently, characterizing the full repertoire of metastable states is fundamental, not only for elucidating biological mechanisms but also for advancing protein design from static scaffolds to dynamic functional systems.

Accurate experimental structures are fundamental to understanding protein function; however, the three primary experimental techniques—X-ray crystallography, cryogenic electron microscopy (cryo-EM), and nuclear magnetic resonance (NMR)—possess inherent limitations in capturing the full spectrum of conformational states. X-ray crystallography and cryo-EM typically capture proteins in kinetically trapped, crystalline, or frozen states. Consequently, the resulting density maps often represent an ensemble average, and the final structural models generally reflect only a single, dominant conformation [2, 3]. This approach fails to reveal the dynamic equilibrium and conformational transitions of proteins under physiological conditions [4, 5]. While NMR spectroscopy can resolve protein dynamics and polymorphism in solution, its application is often constrained by molecular weight, which poses challenges for large macromolecular complexes [6, 7]. Furthermore, NMR-derived structures constitute only a small fraction of the Protein Data Bank (PDB) [8]. As a result, the PDB is predominantly populated by stable, ground-state structures, leading to a significant underrepresentation of functionally essential alternative conformations.

Computational methods have emerged as a powerful alternative; however, they face significant challenges in characterizing the full repertoire of protein states. While classical techniques such as Molecular Dynamics (MD) simulations provide atomic-resolution insights, they remain computationally prohibitive for long-timescale events associated with major conformational transitions and are often limited by force-field inaccuracies [9]. In the deep learning era, architectures such as AlphaFold2 [10] and RoseTTAFold [11] have transformed monomer structure prediction, with AlphaFold-Multimer [12], AlphaFold3 [13] and RoseTTAFold All-Atom [14] extending these capabilities to macromolecular complexes. Despite these advances, leading models like AlphaFold2 are primarily optimized to predict a single, dominant conformation and frequently fail to capture functionally relevant alternative states [15, 16]. To mitigate this, heuristic approaches—including MSA subsampling [15] and clustering [17]—have been developed to isolate distinct evolutionary signals. Nevertheless, their heavy reliance on deep and diverse MSAs significantly constrains their generalizability across the proteome.

Recently, generative models leveraging diffusion [18] or flow matching [19]—such as EigenFold [20], Str2Str [21], and AlphaFlow [22]—have emerged for generating

conformational ensembles. Despite their potential, these continuous-space models are hindered by two interconnected limitations: inadequate conformational exploration and unreliable structural ranking. The first limitation stems from the reliance on L_2 loss functions, which bias models toward minor local perturbations at the expense of large-scale conformational transitions [23]. Consequently, this results in mode averaging on multimodal data, yielding physically implausible structures trapped in high-energy regions rather than distinct metastable basins. The second limitation arises because standard confidence metrics, such as pLDDT and pTM, lack the sensitivity to distinguish valid alternative conformations from noise within large generated ensembles [24, 25]. This highlights a critical gap: the absence of a unified framework capable of globally sampling conformational space to recover accurate alternative states while reliably ranking them.

The emergence of large-scale Protein Language Models (PLMs) marks a paradigm shift from continuous coordinate space to discrete latent representations [26]. Whereas early architectures like ESM2 [27] prioritized sequence information, recent work has increasingly focused on capturing a joint distribution over both sequence and structure. This includes structure-aware PLMs such as SaProt [28], which tokenizes structural backbone coordinates using FoldSeek [29], as well as more recent multimodal generative models like ESM3 [30]. Specifically, ESM3 employs a VQ-VAE [31] to discretize continuous 3D coordinates into a high-dimensional structure token space, implicitly constructing a model-derived energy landscape. However, navigating this landscape remains a significant challenge. The standard sampling strategies in ESM3—namely argmax and iterative decoding—function primarily as local search heuristics comparable to greedy or single-pass Gibbs sampling. Consequently, these methods are often unable to traverse high energy barriers within multimodal landscapes, resulting in a failure to identify distinct conformational basins outside the dominant state.

To address the challenge of limited conformational exploration, we propose MSFold. This framework reconceptualizes the discrete structure token space of ESM3 as an implicit energy landscape amenable to Parallel Tempering (PT) sampling [32], also known as Replica Exchange Monte Carlo. Inspired by the efficacy of Replica Exchange Molecular Dynamics (REMD) [33] in navigating complex continuous landscapes, we adapt this algorithm to the discrete domain of PLMs. By simulating multiple replicas across a spectrum of temperatures, MSFold facilitates efficient global exploration: high-temperature replicas traverse barriers to escape kinetic traps, whereas low-temperature replicas refine local minima. Crucially, the periodic exchange of configurations enables the system to sample the multimodal landscape defined by the PLM without requiring model retraining or fine-tuning. This zero-shot capability directly circumvents the limitations of prior discrete-space methods, such as ESMDiff [23], which necessitate additional training.

To address the concurrent challenge of unreliable structural ranking, we introduce Sequence Log-Likelihood (SLL), a metric quantifying structure-sequence consistency. Although our sampling framework generates a diverse candidate pool, distinguishing physically plausible conformations from artifacts remains difficult using standard metrics. By leveraging the joint distribution learned by ESM3, we utilize its inverse folding capability to compute the probability of the native sequence conditioned on

a generated structure, $P(\text{seq}|\text{stru})$. We posit that SLL serves as a robust proxy for structural quality, effectively differentiating physically realistic states compliant with the native sequence's evolutionary constraints from stochastic, unphysical decoys.

We summarize the primary contributions of this work as follows. First, we introduce MSFold, a novel framework integrating the ESM3 protein language model with Parallel Tempering to robustly explore the protein conformational space. Second, we demonstrate the superior performance of MSFold on a newly constructed, comprehensive benchmark comprising 313 multi-conformation protein pairs. Adopting the convention established by [24], we designate the conformation preferentially predicted by ESM3 as Fold 1 and the alternative state as Fold 2. MSFold achieves a 51.6% success rate in modeling both native states with high accuracy (TM-score ≥ 0.75), surpassing AlphaFold3 (37.8%) [13] by 13.8 percentage points. Notably, MSFold demonstrates balanced accuracy; its aggregated Fold 2 TM-score of 0.740 substantially outperforms all other methods, including AlphaFold3 (0.700). Third, we propose SLL, a novel confidence metric that exhibits a stronger Spearman's rank correlation with true structural quality than pTM and pLDDT, serving as a robust indicator for selecting reliable alternative conformations.

To address the challenge of limited conformational exploration, we propose MSFold. This framework reconceptualizes the discrete structure token space of ESM3 as an implicit energy landscape amenable to Parallel Tempering (PT) sampling [32]. By simulating multiple replicas across a spectrum of temperatures, MSFold facilitates efficient global exploration without requiring model retraining or fine-tuning. On a comprehensive benchmark of 313 multi-conformation protein pairs, MSFold achieves a 51.6% success rate in modeling both native states with high accuracy (TM-score ≥ 0.75), surpassing AlphaFold3 (37.8%) [13] by 13.8 percentage points. Notably, its aggregated Fold 2 TM-score of 0.740 substantially outperforms AlphaFold3 (0.700) and other established baselines, demonstrating a superior capability to escape kinetic traps and sample the multimodal landscape.

To address the concurrent challenge of unreliable structural ranking, we introduce Sequence Log-Likelihood (SLL), a metric quantifying structure-sequence consistency. By leveraging the joint distribution learned by ESM3, we utilize its inverse folding capability to compute the probability of the native sequence conditioned on a generated structure, $P(\text{seq}|\text{stru})$. We posit that SLL serves as a robust proxy for structural quality, effectively differentiating physically realistic states from stochastic, unphysical decoys. Empirically, SLL slightly outperforms standard metrics such as pTM and pLDDT in identifying high-quality structures, providing a more reliable indicator for characterizing complex multi-state landscapes.

We summarize the primary contributions of this work as follows. First, we introduce MSFold, a novel framework integrating the ESM3 protein language model with Parallel Tempering to robustly explore the protein conformational space. Second, we demonstrate that MSFold significantly improves the modeling of multiple native states, achieving state-of-the-art performance on multi-state benchmarks compared to existing methods, including AlphaFold3. Third, we propose SLL, a novel confidence metric based on inverse folding that facilitates the ranking of large sets of candidate structures.

2 Results

2.1 MSFold accurately predicts multiple conformations

To demonstrate the efficacy of our framework, we evaluated MSFold against ten representative baselines on a comprehensive benchmark of 313 multi-conformation protein pairs. This dataset integrates apo/holo pairs [34], fold-switching proteins [16], and a curated subset from the Protein Data Bank (PDB). We employed the TM-score [35] to quantify structural similarity. Adopting the convention established by [24], we designate the conformation preferentially predicted by the base model as Fold 1 and the alternative state as Fold 2. While we retain this nomenclature, we adapt the assignment criterion to reflect the specific predictive bias of ESM3. Consequently, for each target, we report the TM-scores for Fold 1 and Fold 2 separately.

Table 1: Performance comparison on the multi-conformation benchmark. All columns show the average TM-scores for the two conformations, reported as Fold 1/Fold 2. The highest TM-score for each fold is indicated in bold.

Method	Multi-conformation	Apo-holo	Fold-switch	Aggregated Perf.
MSFold (Our method)	0.820/**0.742**	**0.868**/**0.772**	0.749/**0.652**	0.821/**0.740**
AlphaFold2	0.819/0.681	0.858/0.693	0.821/0.589	0.824/0.677
MSA Subsampling	0.787/0.670	0.832/0.686	0.766/0.643	0.790/0.670
MSA Cluster	0.794/0.698	0.849/0.748	0.794/0.627	0.800/0.699
AlphaFold3	**0.825**/0.706	0.866/0.705	**0.833**/0.625	**0.830**/0.700
ESM3 (Argmax decoding)	0.764/0.682	0.825/0.709	0.677/0.588	0.765/0.679
ESM3 (Iterative decoding)	0.792/0.706	0.836/0.725	0.700/0.608	0.792/0.702
ConfDiff	0.672/0.606	0.744/0.637	0.646/0.549	0.679/0.606
Str2Str	0.693/0.613	0.764/0.655	0.679/0.550	0.700/0.613
ESMDiff	0.783/0.696	0.842/0.734	0.700/0.599	0.784/0.694
EigenFold	0.762/0.663	0.804/0.674	0.695/0.560	0.763/0.657
AlphaFlow	0.786/0.689	0.842/0.698	0.753/0.615	0.790/0.685

Figure 1c and Table 1 demonstrate that MSFold establishes a new state-of-the-art in modeling mulltiple conformations. Whereas baseline methods often fail to capture conformational diversity, MSFold achieves a success rate of 51.6%, substantially outperforming both MSA Cluster [17] (38.1%) and AlphaFold3 [13] (37.8%). This performance gap highlights the superior capability of our framework in navigating the energy landscape to identify distinct, biologically relevant states.

Analysis of average TM-scores reveals that MSFold's advantage stems from its more comprehensive prediction across distinct conformational states. In the relatively straightforward Fold 1 task, MSFold demonstrates highly competitive performance with a TM-score of 0.820, on par with top-tier single-structure models such as AlphaFold3 [13] (0.830) and AlphaFold2 [10] (0.823). However, its true value is showcased in the Fold 2 state, which represents a significantly more demanding modeling

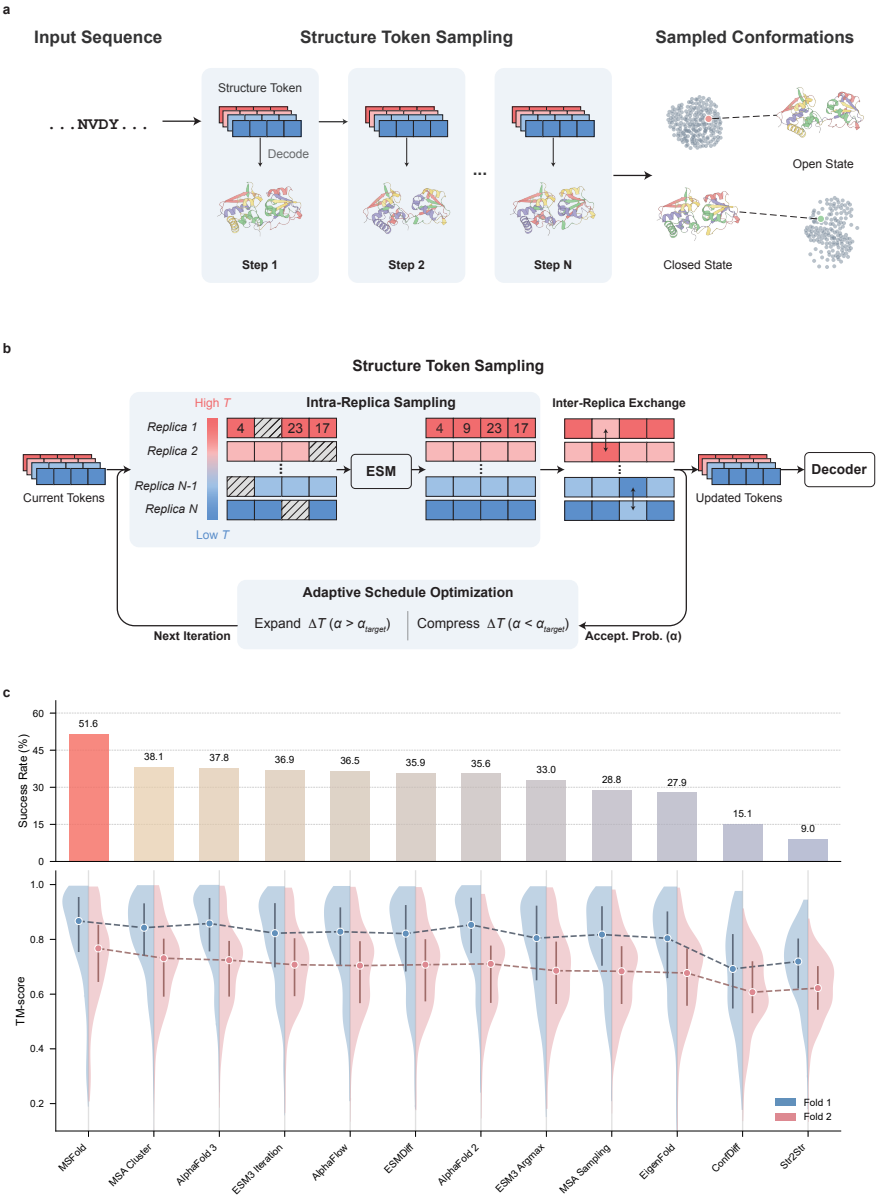

Fig. 1: **The MSFold framework and performance benchmarking. a**, MSFold predicts diverse conformational states from sequences via iterative token sampling. **b**, The Parallel Tempering workflow comprises: (1) *intra-replica sampling* via ESM3; (2) *inter-replica exchange* to overcome energy barriers; and (3) *adaptive schedule optimization* of the temperature ladder. **c**, Performance on 313 multi-conformation pairs. Success rates (TM-score ≥ 0.75 for both states) and TM-score distributions show that MSFold maintains AlphaFold-level accuracy on primary structures (Fold 1) while significantly outperforming existing methods on alternative conformations (Fold 2).

challenge. In this regime, MSFold achieves a TM-score of 0.740, outperforming all other evaluated methods by a substantial margin. The consistent performance drop observed across all models for Fold 2 suggests that this difficulty is likely driven by biases in the training data rather than a specific model limitation. In the prediction of Fold 2, MSFold demonstrates superior sensitivity to the subtle structural signals that define alternative conformations.

A comparison against other ESM3-based techniques highlights the efficacy of advanced algorithmic sampling over structural fine-tuning. In the Fold 1 regime, MSFold achieves a TM-score of 0.821, substantially surpassing the standard ESM3 baselines of argmax (0.765) and iterative decoding (0.792). This improvement is primarily attributed to the use of lower sampling temperatures, which favor high-precision local refinement. Although low-temperature sampling typically constrains conformational diversity—often leading to mode collapse—MSFold circumvents this trade-off via the replica exchange mechanism. This allows the system to retain the precision of low-temperature replicas while simultaneously accessing the broad conformational space explored by high-temperature replicas. Consequently, MSFold achieves a Fold 2 TM-score of 0.740, significantly outperforming standard iterative decoding (0.702). Furthermore, MSFold outperforms ESMDiff [23], a model explicitly fine-tuned on structural data, across both Fold 1 (0.821 vs. 0.784) and Fold 2 (0.740 vs. 0.694). These results provide compelling evidence that applying advanced sampling algorithms to a frozen base model can unlock significant performance gains, effectively surpassing the benefits of computationally expensive retraining when modeling multi-state proteins.

2.2 MSFold demonstrates robust generalization on a strict temporal-holdout dataset

A critical consideration in our benchmark design addresses the potential for data leakage, as the curation of our 313-pair dataset did not enforce a strict temporal cutoff relative to the training corpora of the evaluated baselines. This design choice was deliberate, prioritizing the assembly of a benchmark with sufficient scale and diversity—encompassing fold-switching proteins and ligand-induced conformational changes—to rigorously assess multi-state modeling capabilities. Despite this potential overlap, we maintain that the validity of our core findings remains intact, supported by three key observations.

First, the evaluation objective is not merely to assess the model's ability to recall a single ground-truth structure (whether Fold 1 or Fold 2), but rather to sample a conformational ensemble that spans multiple states. The primary goal is to determine whether our framework empowers the base model to overcome its intrinsic limitations, enabling the generation of a diverse set of conformations from a single sequence that simultaneously captures representatives of distinct metastable basins.

Second, existing methods frequently fail to recover multiple conformational states, even when the target proteins are included in their training data. To quantify this, we analyzed AlphaFold2 as a representative case study, filtering our benchmark based on its training cutoff (April 30, 2018). Even within the subset where both conformations were chronologically available for training, AlphaFold2 achieved a success rate

of merely 0.431. This result indicates that the inability to sample conformational diversity is not a consequence of data scarcity, but rather an intrinsic limitation of the model's architecture and training objective, which bias it toward a single static structure even for known targets.

Third, to rigorously validate our findings, we curated a stringent temporal-holdout subset comprising 32 pairs released after the ESM3 training cutoff (May 1, 2020). We selected this specific date because adopting the latest cutoff among all baseline methods (e.g., June 2023 for Str2Str) would yield an empty test set. This setup establishes a highly conservative evaluation scenario: while this subset remains strictly out-of-distribution for ESM3, it potentially includes training data for models with later cutoffs, such as AlphaFold3 (Sept. 2021) and ConfDiff (Dec. 2021).

On this challenging holdout set, the performance advantage of ESM3 is not only maintained but further amplified (Table 2). ESM3 (Fold 2 TM: 0.759) significantly outperforms all other baselines, including AlphaFold3 (Fold 2 TM: 0.671) and ESMdiff (Fold 2 TM: 0.672). These results demonstrate that ESM3 generalizes robustly to out-of-distribution data, confirming its position as the state-of-the-art method on this stringent temporal-holdout set while remaining unaffected by potential data leakage or memorization.

Table 2: Comparison on the strict temporal-holdout subset. This subset (32 pairs) contains proteins released after the MSFold/ESM3 training cutoff (May 1, 2020). Cutoffs later than this date are highlighted in bold, indicating this set was "seen" during their training.

Method	Training Cutoff	Success Rate	Performance (Fold 1 / Fold 2)
MSFold (Our method)	2020.5.1	**0.625**	**0.822** / **0.759**
AlphaFold2	2018.4.30	0.343	0.775 / 0.606
MSA Subsampling	2018.4.30	0.313	0.723 / 0.609
MSA Cluster	2018.4.30	0.438	0.739 / 0.634
AlphaFold3	**2021.9.30**	0.406	0.792 / 0.671
ESM3 (Argmax)	2020.5.1	0.375	0.735 / 0.676
ESM3 (Iterative)	2020.5.1	0.469	0.764 / 0.692
ConfDiff	**2021.12.31**	0.094	0.599 / 0.553
Str2Str	**2023.6.9**	0.031	0.672 / 0.549
ESMdiff	2020.5.1	0.375	0.744 / 0.672
EigenFold	2020.4.30	0.312	0.699 / 0.609
AlphaFlow	**2023.1.0**	0.344	0.719 / 0.639

2.3 Parallel Tempering sampling effectively uncovers the diverse conformational landscape defined by the PLM

The two default sampling strategies in ESM3 fail to generate diverse conformational ensembles, necessitating a more advanced sampling scheme. Argmax decoding is purely deterministic, seeking only the single most probable structure (the global maximum), which renders it incapable of exploring the conformational space. Furthermore,

iterative decoding, essentially a single pass of Gibbs sampling, functions as a local search method that struggles to traverse high-energy barriers, often trapping the model in suboptimal local minima.

In these sampling algorithms, increasing the temperature introduces structural diversity but inevitably degrades structural accuracy. Governed by the softmax formulation $P(x_i) \propto \exp(z_i/T)$, a higher temperature T flattens the logit distribution, thereby increasing the probability of sampling diverse yet lower-confidence candidates. To quantify these effects, we tracked the TM-scores of sampled structures relative to both native states throughout the process. Using the LAO-binding protein (LAOBP)—which features distinct Open (2LAO) and Closed (1LAH) conformations—as a representative case, Figure 2 illustrates an inherent trade-off where temperature adjustments enhance either diversity or accuracy at the expense of the other. At low temperatures (Figure 2b, d), the model favors high-probability tokens, causing sampled structures to converge strictly toward the Closed (1LAH) state. This greedy strategy restricts exploration to a single conformational basin, preventing the discovery of the alternative Open state (2LAO). Conversely, raising the temperature to encourage exploration (Figure 2a, c) precipitates a sharp decline in structural quality. In this regime, sampled structures exhibit excessive stochasticity and lack physical plausibility, often failing to maintain secondary structures.

In contrast, the MSFold successfully captures both native states. As illustrated in Figure 3a, MSFold generates a structural ensemble that spans the conformational landscape, with individual samples converging to distinct native basins. The sampling trajectory reveals transitions between these states, confirming the ability of MSFold to escape local energy minima and explore the broader configuration space. Notably, the trajectory alternates between conformations, demonstrating that the method effectively overcomes the sampling barrier for Fold 2, a region inaccessible to standard ESM3 strategies (Figure 2e).

Visual inspection of the sampled structures reveals the emergence of two distinct clusters, each corresponding strongly to one of the native states, connected by discernible transition paths. To rigorously analyze this landscape, we projected the sampling trajectory onto a 2D manifold by computing a pairwise similarity matrix of the sampled structures. Using $1 -$ TM-score as the distance metric, we applied Uniform Manifold Approximation and Projection (UMAP) [36] for dimensionality reduction, embedding the known native states as reference points. As shown in Figure 3b, the resulting UMAP projection exhibits clear structural clustering. Superimposing the transition trajectory from Figure 3a onto this landscape reveals a coherent sampling path bridging the two clusters. This confirms that the observed state transitions are not stochastic fluctuations but directed movements along a functionally relevant pathway, thereby validating the efficacy of our sampling strategy.

Comparable performance is observed for Adenylate Kinase (Figure S2) and the catalytic antibody 48G7 Fab (Figure S1). Mainly in that MSFold successfully captures both native states and exhibits robust transitions between them.

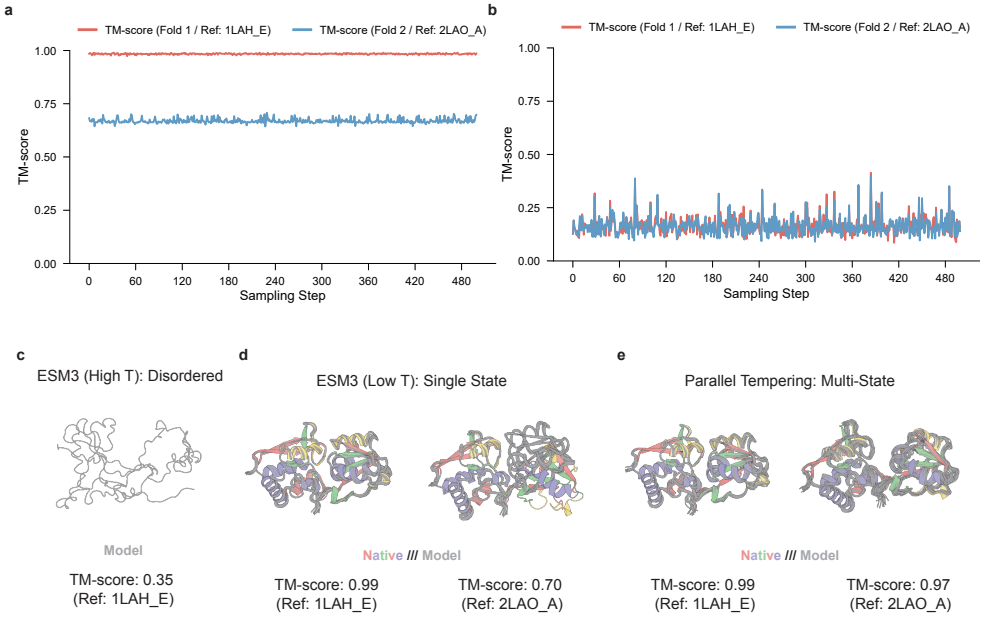

Fig. 2: **Comparison of sampling behaviors between standard ESM3 iterative decoding and MSFold. a, d**, Limitations of low-temperature sampling. **a**, The TM-score trajectory of ESM3 iterative decoding at low temperature shows the model getting trapped in a single local minimum corresponding to the Closed state (1LAH_E), consistently failing to sample the Open state (2LAO_A). **d**, Structural superposition of 10 generated models (grey) confirms this mode collapse, as all predictions align exclusively with the Closed native reference while missing the Open state. **b, c**, Limitations of high-temperature sampling. **b**, Increasing the temperature to encourage diversity results in a sharp decline in structural quality, with TM-scores dropping below 0.4 for both states. **c**, A representative structure from this regime exhibits global disorder and a loss of coherent secondary structure, illustrating the trade-off between exploration and physical plausibility. **e**, Efficacy of the MSFold framework. In contrast to standard decoding, MSFold successfully navigates the energy landscape to sample both conformational states. The superposition of 10 sampled structures (grey) demonstrates high-accuracy coverage of both the Closed (1LAH_E) and Open (2LAO_A) native conformations within a single sampling run.

2.4 Distinct Exploratory Roles of Replicas at Different Temperatures

To characterize the sampling behavior of replicas across varying temperatures, we visualized the generated conformational ensembles via UMAP projections (Figure 4c–e). These projections reveal distinct temperature-dependent sampling dynamics throughout the ensemble.

The UMAP projections highlight a fundamental disparity in the sampling behaviors of high- and low-temperature replicas. High-temperature sampling coalesces

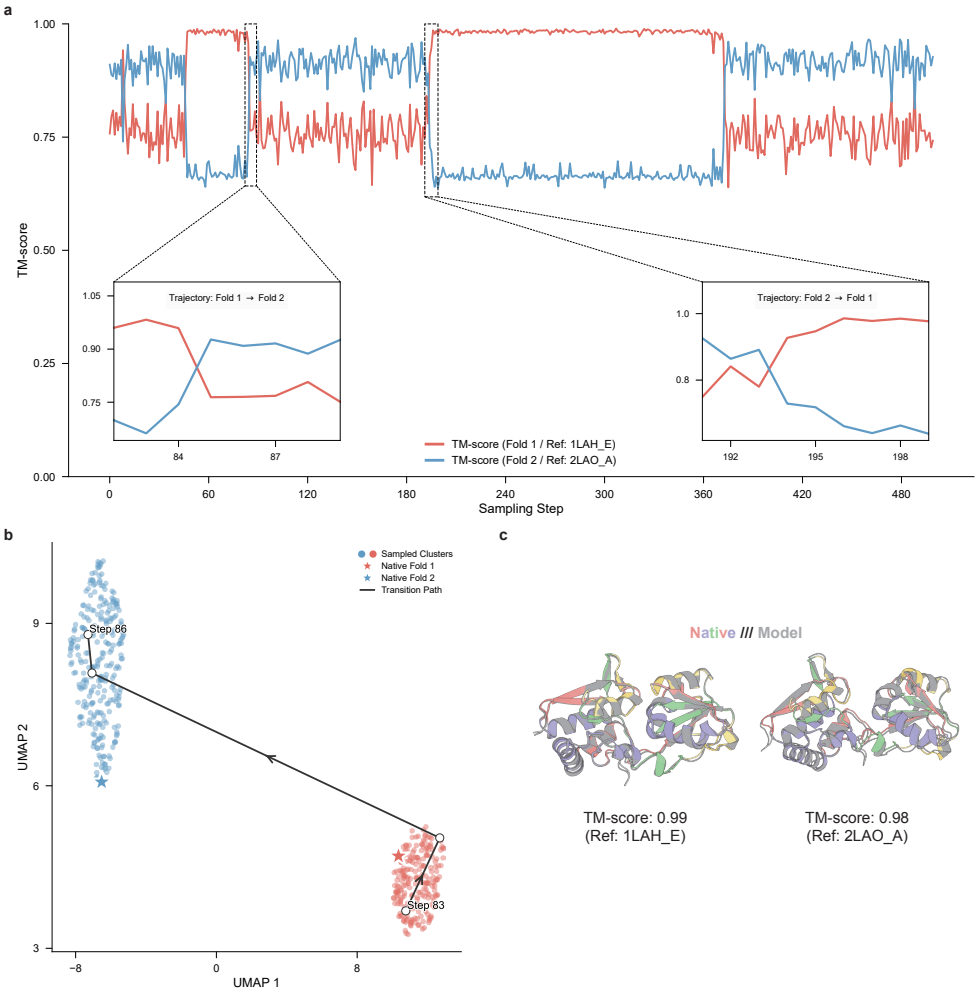

Fig. 3: **MSFold captures both native states of the LAO binding protein.** **a**, Evolution of structural similarity (TM-score) relative to the closed state (Fold 1, PDB: 1LAH_E; red) and the open state (Fold 2, PDB: 2LAO_A; blue). The trajectory consistently samples high-accuracy structures of either the closed or open state, rather than intermediate mixtures. Two insets highlight the reversible transitions (round-trips) between these conformations during the sampling process. **b**, UMAP projection of the sampled conformational space. Each point represents a sampled structure, revealing two clusters corresponding to the native references (stars). The arrowed path tracks the specific transition event shown in the left inset of **a** (Steps 83–86). **c**, Structural superpositions of predicted models (grey) onto native references (colored). The predicted structures exhibit high structural similarity to the native states, achieving TM-scores of 0.99 and 0.98, respectively.

Fig. 4: **Token diversity, transition analysis, and multi-replica conformational space. a**, Residue-wise token distribution and entropy (residues 60–120), highlighting position-specific variability in token diversity. **b**, Radial plots comparing token identity during transitions and against native references. High structural similarity is maintained despite low sequence identity in the structure token space. **c**–**e**, UMAP visualizations of conformational space across temperature replicas. As temperature decreases (Replica 11 to 37), diffuse, merged clusters resolve into two distinct, compact states (Fold 1, red; Fold 2, blue) centered on native references (stars).

conformational clusters into a single, continuous landscape, indicative of diffuse exploration. Functioning as global explorers, these replicas possess sufficient thermal energy to surmount high energy barriers, thereby traversing the broad conformational space and accessing disconnected energy basins. Conversely, low-temperature sampling yields discrete, compact clusters, characteristic of highly localized confinement. These replicas facilitate local refinement, remaining confined within specific energy minima to extensively sample the local phase space. This localization enables the generation of high-fidelity structures that accurately characterize stable states. Although low-temperature replicas are kinetically trapped in local minima, their synergy with the global exploration provided by high-temperature replicas underpins the efficacy of Parallel Tempering.

As shown in Figure 4d,e, temperature significantly alters the relative populations of the two conformational clusters. This temperature-dependent shift demonstrates that replicas sample available states from distinct probability distributions. Consequently, distributing multiple replicas across a broad temperature range improves sampling diversity. This approach increases the probability of capturing rare or energetically prohibitive states often missed by single-temperature simulations.

2.5 Token Space Dissimilarity Does Not Preclude Structural Similarity

Rapid transitions between the two conformational states occur within a few sampling steps (Figure 3a,b). To investigate the mechanism driving these transitions, we analyzed the structure tokens of the conformations immediately flanking a representative jump (Figure 4b). The discrete representations shift significantly, with 72.3% of residues altering their assigned structure tokens. Comparing these jump-related conformations to the native states reveals an apparent paradox: despite highly dissimilar structure tokens, the 3D structural fidelity remains high. For example, although the structure tokens of step 84 and the 1LAH state differ by 88.7%, their TM-score is 0.96. These results demonstrate a clear decoupling between discrete token identity and 3D structural conformation.

This apparent paradox arises because the ESM3 decoder reconstructs 3D structures from continuous embeddings rather than raw token IDs. Consequently, embedding-space similarity dictates structural proximity. Because ESM3 employs a VQ-VAE [31] to discretize 3D protein structures, each predicted conformation possesses three distinct representations: 3D coordinates (structure space), continuous latent vectors (embedding space), and discrete indices (token space). That vastly different token sequences yield nearly identical 3D structures indicates that the token-to-structure mapping is not distance-preserving. To investigate this phenomenon, we analyzed the correlations and marginal distributions across three metrics: token identity, embedding similarity, and structural similarity (TM-score).

Figure 5 illustrates the pairwise correlations among the similarities across these three spaces. We defined token identity as $1 -$ normalized Hamming distance, embedding similarity as $1 -$ cosine distance, and structural similarity as the TM-score. The distributions of these metrics differ significantly: token identity is confined to a narrow range ($[0, 0.4]$). In contrast, both embedding and structural similarities exhibit

broader distributions across the full $[0, 1]$ interval, displaying similar bimodal peaks near the extremes. Furthermore, although the correlation coefficient between token identity and TM-score is relatively high (0.84 to 0.89), the Root Mean Square Error (RMSE) remains notably large (0.54). This high RMSE quantitatively confirms that the token space is not distance-preserving with respect to the structural space; dissimilar token sequences frequently correspond to highly similar structures. Conversely, embedding similarity aligns more strongly with structural similarity, yielding a higher correlation (0.89 to 0.96) and a significantly lower RMSE (0.11). These results demonstrate that although discrete token IDs in ESM3 do not accurately reflect physical structural distances, their associated continuous embeddings maintain a high degree of isometry with the 3D structural landscape.

2.6 Heterogeneity in Residue-wise Structure Token Entropy

Figure 4a illustrates the distribution and entropy profiles of the sampled structure tokens. Structural token diversity varies significantly among individual residues. Although certain regions maintain highly conserved, near-deterministic token profiles, others exhibit significant entropy, reflecting broader ensembles of potential local geometries. This variance indicates that the model's structural assignment confidence is not uniform; instead, it remains highly sensitive to local residue context.

These high-entropy sites indicate that the learned structural representation captures specific residues with less constrained or highly variable local geometries. Furthermore, these residues represent primary degrees of freedom within the latent space, functioning as structural pivot points that facilitate state transitions. For MSFold , this non-uniform entropy profile allows the model to strategically explore regions of high structural uncertainty. Focusing exploration on these flexible segments while preserving stable, low-entropy regions enables the method to navigate the global conformational space and uncover otherwise inaccessible alternative states.

2.7 Ranking Predicted Structural Models via Sequence-Structure Consistency

We evaluated the efficacy of three confidence metrics—Sequence Log-Likelihood (SLL), pTM, and pLDDT—for selecting high-quality conformations from the MSFold pool across the multi-conformation dataset. After independently ranking all candidate structures by each metric, we identified the maximum TM-score among the top k candidates ($k = 1, 5, 10, 20$). Table 3 demonstrates that SLL marginally outperforms the alternative metrics in structural selection. Specifically, SLL consistently yielded the highest TM-scores across all top-k cutoffs for both Fold 1 and Fold 2, slightly exceeding pTM and pLDDT in retrieving high-quality structures. Nevertheless, a substantial performance gap persists between the top-k TM-scores and the theoretical upper bound (0.821/0.740), which represents the best conformation available within the entire pool. This discrepancy illustrates the ongoing challenge of reliably identifying the optimal conformation using current confidence metrics.

To further investigate this relationship, we analyzed a multi-conformation protein (1YNL_L/2CGR_L) by correlating confidence metrics with actual TM-scores.

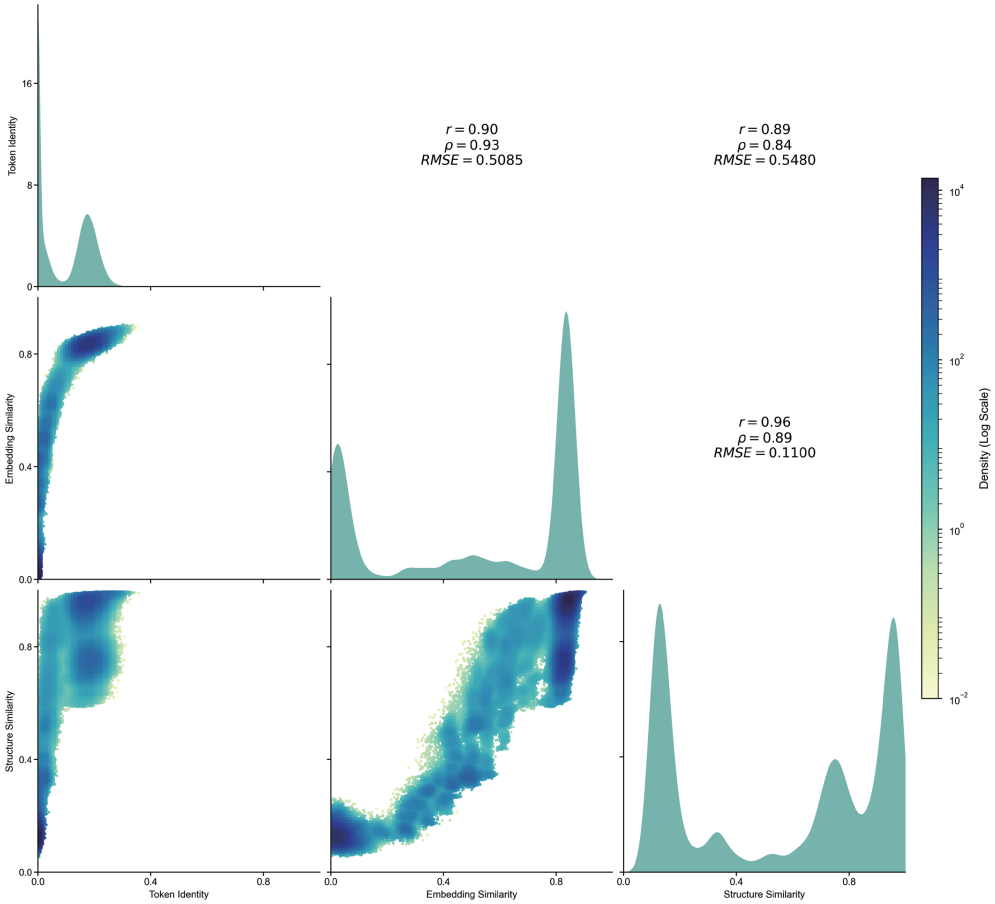

Fig. 5: **Consistency analysis of pairwise similarities across token, embedding, and structure spaces.** This figure analyzes the mapping properties from the physical structure space to the ESM3 latent spaces (discrete tokens and continuous embeddings). **Diagonal panels**: Marginal probability distributions of pairwise similarities for Token Identity, Embedding Similarity, and Structure Similarity. **Lower and Upper triangles**: The lower triangle displays 2D log-density histograms of joint distributions, while the upper triangle reports corresponding statistical correlations and RMSE. **Key Insight**: The discrete token space exhibits a *non-isometric* nature, characterized by low token identity even among structurally similar pairs (bottom-left). In contrast, the continuous embedding space maintains a distance-preserving relationship with the physical structure space (bottom-center), showing high consistency.

Derived from sequence-structure consistency, SLL correlates strongly with the TM-score, effectively characterizing structural quality (Figure 6a). Furthermore, although pLDDT and pTM rankings deviate significantly from actual structural quality, the SLL ranking aligns closely with the ground truth (Figure 6b). These results demonstrate that SLL provides a more accurate assessment of structural integrity in these challenging cases.

Overall, SLL performs comparably to pTM and pLDDT, exhibiting a slight advantage in structural selection (Table 3). Crucially, SLL provides a novel confidence metric derived from sequence-structure consistency. Rather than replacing established metrics, it serves as a complementary indicator to comprehensively evaluate structural confidence in multi-conformation proteins. Consequently, integrating SLL with existing scoring functions can more accurately characterize structural quality and facilitate the selection of reliable conformations.

Table 3: Quality of structures selected by different confidence metrics.

Metric	Ideal TM-score	Maximum TM-score Achieved in Top k Candidates (Fold 1 / Fold 2)			
		Top 1	Top 5	Top 10	Top 20
SLL	0.821/0.740	**0.720/0.611**	**0.749/0.639**	**0.757/0.648**	**0.766/0.657**
pTM	-	0.715/0.607	0.743/0.635	0.755/0.644	0.762/0.652
pLDDT	-	0.697/0.593	0.731/0.627	0.743/0.639	0.752/0.648

2.8 Inter-replica Exchange as the Critical Determinant of Sampling Performance

To evaluate the contribution of each algorithmic component, we performed an ablation study using 10 representative multi-conformation protein pairs. We assessed two primary ablations: *No exchange*, which disables the inter-replica configuration exchange—the core of Parallel Tempering; and *No block*, which replaces the block Gibbs update with single-residue sampling. We compared these configurations against ESM3's default iterative and argmax decoding strategies to establish a generative performance baseline.

Table 4a identifies replica exchange as the most critical factor for successful multi-state modeling. Disabling this mechanism (*No exchange*) severely degrades performance; Fold 1 scores drop from 0.840 to 0.802, and Fold 2 scores decrease from 0.792 to 0.775. This decline indicates that lacking the global exploration facilitated by configuration swaps, the model remains trapped in local minima, failing to cross the high-energy barriers separating distinct conformations. Furthermore, the *No block* configuration yields only a marginal performance decrease, indicating that although block-wise updates improve local sampling efficiency, they remain secondary to the global search capacity of Parallel Tempering. Results from the ablation experiments and comparative analysis with ESM3 reveal that the generative potential of protein

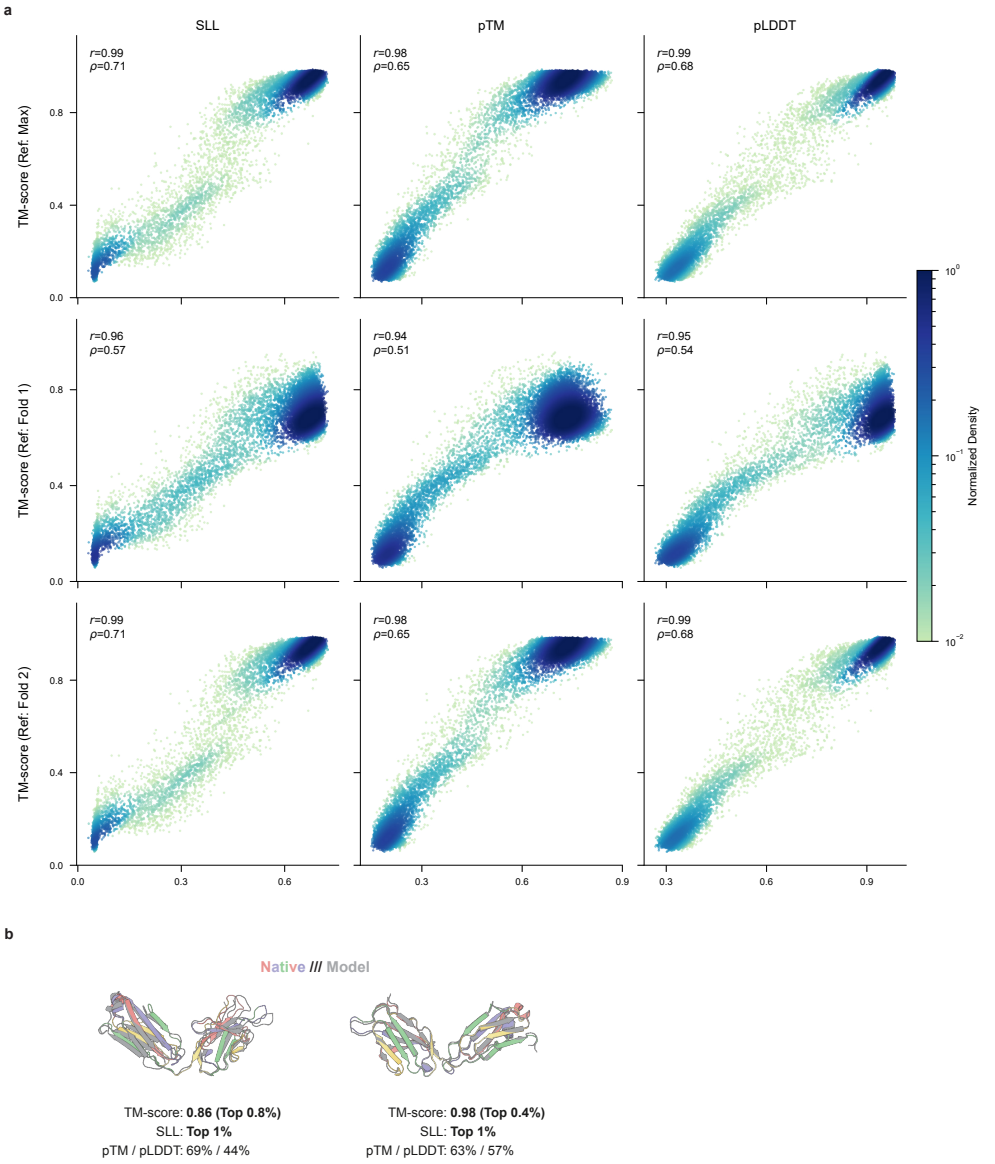

Fig. 6: **Evaluation of Sequence Log-Likelihood (SLL) against standard confidence metrics for structure ranking. a**, Correlation analysis between confidence scores and structural accuracy. SLL exhibits a strong positive correlation with TM-score, validating its utility as a confidence metric comparable to established standards like pTM and pLDDT. **b**, Case study highlighting the sensitivity of SLL in ranking alternative conformations. For the two high-fidelity structures shown, standard metrics fail to recognize their quality, assigning them low rankings. In sharp contrast, SLL correctly identifies both as top-tier candidates (Top 1% of the pool), demonstrating its ability to retrieve valid physical structures that are misclassified by conventional scores.

language models is fundamentally constrained by the sampling algorithm used to traverse the latent space.

A parameter sensitivity analysis of the replica count and sampling steps establishes the practical limits of our method (Table 4b,c). Both parameters exhibit clear diminishing returns. Although increasing the number of replicas initially improves accuracy by broadening coverage of the energy surface, performance saturates beyond 30 replicas. Additional replicas likely introduce sampling stochasticity rather than genuine gains in search capacity. Similarly, extending the sampling steps yields progressively fewer marginal benefits beyond a specific threshold. These findings inform resource optimization, indicating that moderate computational configurations achieve high-fidelity sampling while preventing the linear memory and latency overheads associated with excessive replicas or steps.

Finally, we evaluated the computational feasibility of the MSFold method. Using a single NVIDIA A800 GPU, processing a 200-residue protein requires approximately 20 GiB of memory and 8 hours of runtime. For smaller 50-residue proteins, the runtime drops significantly to 0.6 hours. Furthermore, decreasing the replica count and sampling steps can further reduce inference time without substantially compromising structural accuracy or conformational diversity.

Table 4: **Ablation Study and Parameter Sensitivity Analysis.**

(a) Sampling Strategies			**(b) Number of Replicas**			**(c) Sampling Steps**		
Method	Fold 1	Fold 2	Replicas	Fold 1	Fold 2	Steps	Fold 1	Fold 2
MSFold (Ours)	**0.840**	**0.792**	20	0.831	0.782	100	0.825	0.778
No block	0.823	0.792	30	0.839	0.794	200	0.834	0.781
No exchange	0.802	0.775	40	0.840	0.792	300	0.836	0.787
ESM3 (Iterative)	0.794	0.747	50	0.837	0.798	400	0.836	0.788
ESM3 (Argmax)	0.792	0.720	–	–	–	500	**0.840**	**0.792**

2.9 Benchmark construction and experimental setup

To establish a comprehensive baseline, MSFold was compared against a diverse set of generative and predictive methods. For AlphaFold2 [10], 25 structures were generated across its five internal models. For MSA Subsampling [15], 10 structures were obtained via 10 independent executions (`max_msa_clusters` = 32). MSA Cluster [17] was configured with `min_eps` = 3. For ConfDiff [37] and Str2Str [21], 200 structures each were sampled under default configurations. EigenFold [20] and AlphaFlow [22] were used to generate 5 and 10 structures, respectively. For AlphaFold3 [13], 25 structures were produced across five random seeds. Finally, the 1.4B-parameter ESM3 [30] generated 100 structures via both argmax and iterative decoding, while 100 structures were sampled from ESMdiff [23] using 25 steps.

The benchmark dataset comprises three subsets: apo/holo pairs [34], fold-switching proteins [16], and a custom-curated subset from the Protein Data Bank (PDB). The curation pipeline for the PDB subset proceeded as follows: structures were initially

filtered by length (50–500 residues) and coverage (> 95%). Next, MMseqs2 [38] clustered the sequences at 100% identity. Within these clusters, US-align [39] identified conformationally diverse pairs (TM-score < 0.8). Finally, DSSP [40, 41] was utilized to exclude disordered proteins. To ensure consistency, the apo/holo and fold-switch datasets were subjected to the same filtering criteria, specifically removing entries where the sequence length fell outside the 50–500 range, where the conformational TM-score was above 0.8, or where the sequence identity was below 95%. Integrating these sources yielded a comprehensive benchmark of 313 multi-conformation protein pairs. The statistical distributions of sequence lengths and conformational diversity across these 313 targets are summarized in Figure 7.

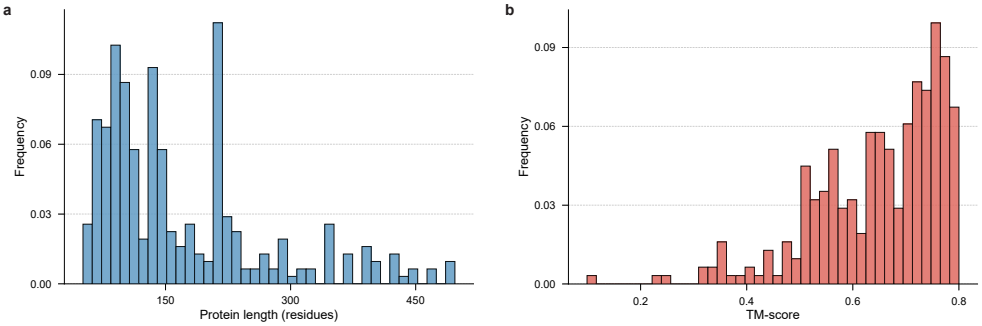

Fig. 7: **Statistics of the multi-conformation benchmark dataset. a**, Distribution of protein sequence lengths for the 313 selected targets. **b**, Distribution of pairwise TM-scores calculated between the two distinct native conformations (Fold 1 and Fold 2) for each target.

2.10 Inertia of the discretized structural representation limits the impact of local logit perturbations

To evaluate the preformance in predicting alternative conformations, we curated a test set of 20 protein pairs exhibiting significant conformational diversity (average TM-score of 0.520 between native conformers). We specifically focused on targets where the model initially showed a large discrepancy between the Fold 1 and Fold 2. To encourage the sampling of Fold 2, we investigated guided sampling strategies by manipulating structural token logits—specifically, suppressing the logits of the known conformation's tokens (Token-level) or their associated cluster centers (Cluster-level).

Table 5 summarizes the performance of these strategies. Contrary to the hypothesis that penalizing the ground truth would drive the model toward alternative minima, we observed negligible structural changes. The Token-level strategy yielded almost no deviation from the original predictions. Similarly, cluster-level strategies resulted in slight structural variation despite suppressing broader regions of the token space; even the most aggressive strategy (Cluster 4) improved the TM-score by only 0.005.

These results indicate that large perturbations in the logit space do not translate into significant geometric transitions in the decoded structure.

Table 5: Performance of Guided Sampling Strategies for Alternative Conformation Prediction

Method	Fold 1 (TM-score)	Fold 2 (TM-score)
Cluster 1	0.913	0.615
Cluster 2	0.826	0.613
Cluster 3	0.916	0.619
Cluster 4	0.913	**0.628**
Cluster 5	0.910	0.614
Cluster 6	0.892	0.615
Cluster 7	0.915	0.618
Cluster 8	0.910	0.619
Cluster 9	0.859	0.613
Cluster 10	0.890	0.613
Token-level	**0.919**	0.619
MSFold	**0.919**	0.623

The limited efficacy of logit-based strategies in inducing structural changes—despite significant modifications to the sampling distribution—attributes this observed structural inertia to intrinsic properties of the VQ-VAE discretization and decoding mechanism. We detail this analysis further in Section 2.11.

2.11 Mechanisms Underlying the Non-linear Mapping Between Discrete Token Semantics and 3D Conformational Outputs

Integrating a Vector Quantized Variational Autoencoder (VQ-VAE) into ESM3 to discretize protein structures establishes a complex mapping between discrete tokens and their corresponding 3D geometries. Our results demonstrate two counterintuitive phenomena: highly dissimilar structure tokens can produce nearly identical structures, and directly manipulating token logits during sampling rarely alters the final conformation. Explaining these observations requires analyzing the semantic structure of the codebook alongside the operational logic of the decoder.

Significant semantic redundancy within the learned codebook drives this behavior. Recent analyses [42][43] of protein structure tokenization reveal that ESM3-learned tokens cluster densely in the latent embedding space. This clustering arises because valid protein conformations—primarily variations of α-helices and β-sheets—occupy a low-dimensional manifold. Consequently, many discrete tokens map to highly proximal embedding representations, effectively functioning as semantic synonyms. Because of this redundancy, the sampling algorithm can substitute tokens within the same cluster, yet this substitution does not alter the structural information provided to the

decoder. Thus, even if logit modification suppresses specific token classes, the sampler likely selects a semantically equivalent neighbor; this yields a different token index but preserves the final 3D structure.

This redundancy is further complicated by a fundamental disconnect between the logit space and the resulting geometric changes. In continuous latent space models, such as standard VAEs or GANs, moving along a latent gradient typically yields smooth and predictable semantic transitions. However, the discrete latent space of a VQ-VAE creates a highly non-linear relationship where the magnitude of logit perturbation does not correlate predictably with geometric deviation. Large perturbations intended to weaken a known conformation in the logit space do not ensure that the generated structure will diverge significantly. This unpredictability suggests that structural generation cannot be reliably controlled through simple logit-level guidance, as the mapping from discrete indices to 3D coordinates lacks a linear, distance-preserving property.

Ultimately, the structure decoder functions as a robust implicit prior, frequently overriding local discrete signals. Tokens do not dictate rigid structural instructions; instead, they provide soft contextual cues interpreted alongside the amino acid sequence and global physical constraints. Because the Transformer architecture relies heavily on broad contextual dependencies, the decoder treats local token perturbations as noise. It resolves local inconsistencies by integrating the surrounding unperturbed context and enforcing learned physical laws. Although this dominant prior ensures structural plausibility, it simultaneously dampens the effect of local modifications. Consequently, the final 3D conformation remains highly resistant to isolated alterations in the token sequence.

3 Discussion

The efficacy of MSFold arises from a conceptual shift: reinterpreting the learned structure token space of the Protein Language Model (PLM) not merely as a sequence of discrete labels, but as an implicit energy landscape amenable to physical sampling. Our results indicate that ESM3's default decoding tends to collapse into a single dominant state, failing to produce the diverse ensembles required for multi-state modeling. In contrast, the Parallel Tempering (PT) mechanism facilitates barrier crossing, enabling directed transitions between distinct conformational clusters—a capability absent in standard iterative decoding. Furthermore, block-based Gibbs sampling significantly enhances sampling efficiency by capturing synergistic residue interactions, thereby enabling the model to navigate the local energy landscape more effectively than single-site updates. In the inter-replica exchange phase, the strategy of swapping local configurations—complemented by an adaptive temperature schedule—maximizes exchange efficiency. This approach effectively overcomes the sampling bottlenecks common in high-dimensional state spaces, where global swaps often suffer from prohibitively low acceptance probabilities due to insufficient configurational overlap. These findings validate the integration of classical statistical physics algorithms with deep generative models to access the latent structural diversity of PLMs.

To rank the predicted structures, we introduce the Sequence Log-Likelihood (SLL), an inverse-folding metric that evaluates structural plausibility based on consistency with the native sequence. The fact that SLL slightly outperforms pLDDT and pTM further underscores its reliability. Given that SLL assesses structural plausibility from the perspective of sequence recovery, it offers a distinct viewpoint compared to traditional metrics. Therefore, integrating SLL with the widely used pTM and pLDDT provides a more stable and comprehensive confidence indicator for evaluating generated conformational ensembles.

However, the framework operates within specific boundaries. As a sampling strategy, MSFold explores—but does not define—the energy landscape; performance is therefore upper-bounded by the quality of the underlying ESM3 model and its VQ-VAE tokenizer. This dependence imposes two constraints: first, if the base model's landscape lacks a stable minimum for a specific conformation, sampling cannot recover it; second, the discrete approximation of continuous conformational space introduces representational bottlenecks. Reconstruction errors from the VQ-VAE decoder establish a hard ceiling on structural fidelity. Future improvements will necessitate not only advanced sampling algorithms but also the co-development of higher-resolution tokenizers to minimize these discretization artifacts.

A requisite trade-off for this exhaustive exploration is increased computational cost relative to single-pass inference. Simulating multiple replicas scales resource consumption linearly. However, our ablation studies demonstrate that this cost is tunable; reducing the number of replicas or sampling steps decreases memory and time requirements with strictly bounded performance degradation. This flexibility allows MSFold to adapt to varying regimes, from high-fidelity structural characterization to scenarios demanding rapid generation.

In conclusion, MSFold provides a robust framework for overcoming the local sampling limitations of protein language models. By enabling the exploration of multi-modal landscapes without retraining, it captures challenging conformational changes more effectively than existing methods. Ultimately, our results highlight that physics-informed generative sampling frameworks serve as a key to unlocking the latent generative capabilities of PLMs, effectively bridging the gap between static predictions and multi-state conformational ensembles.

4 Method

4.1 Multi-conformation Generation using ESM3 and Parallel Tempering

Our conformational sampling framework builds upon the ESM3 generative model [30], which jointly models protein sequence and structure. Crucially, ESM3 establishes a probabilistic link via the conditional distribution $P(\text{Sequence} \mid \text{Structure})$. This formulation allows us to evaluate $\log P(\text{Structure} \mid \text{Sequence})$, which serves as an implicit, model-derived pseudo-energy function. This pseudo-energy defines a complex conformational energy surface, which our approach aims to efficiently explore.

Furthermore, ESM3 inherently compresses the continuous 3D coordinate space into a discrete vocabulary of structure tokens. Direct sampling in continuous 3D

space—as seen in traditional Molecular Dynamics—is computationally prohibitive and frequently suffers from poor mixing. By operating within this discrete token space instead, we reframe the continuous sampling challenge as a tractable, discrete combinatorial problem. Finally, the ESM3 decoder provides a fast, deterministic mapping from any given sequence of structure tokens back to full 3D atomic coordinates.

Although tokenization reframes the sampling challenge, it does not inherently resolve the difficulty of global exploration. The resulting discrete search space remains exceptionally large: because ESM3 utilizes a vocabulary of 4096 distinct structure tokens, a protein of length L yields a vast combinatorial space of 4096^L. Much like its continuous counterpart, this discrete energy surface is characterized by deep local minima—representing distinct conformational states—separated by high energy barriers. Because ESM3's default sampling strategies function primarily as local search techniques, they are constrained to identifying the single most probable conformation. Consequently, these baseline methods fail to cross high energy barriers to discover alternative conformational basins.

To address this fundamental limitation, we employ Parallel Tempering (PT) [32], a technique designed to effectively sample multi-modal distributions characterized by high energy barriers (Figure 1). PT simulates multiple replicas of the system in parallel, maintaining each at a distinct temperature. High-temperature replicas (T_{high}) experience a flattened energy surface, enabling them to easily traverse energy barriers and perform global exploration. Concurrently, low-temperature replicas (T_{low}) operate on the unmodified energy surface, performing fine-grained local refinement within deep energy basins. The core mechanism of PT—configuration exchange between replicas at adjacent temperatures—allows the low-temperature system to escape kinetic traps by inheriting globally explored states from the high-temperature replicas. Consequently, this approach ensures comprehensive and efficient sampling of the conformational space defined by the ESM3 model.

Algorithm 1 outlines the overall sampling procedure. Initially, the protocol establishes all replica states and temperatures. The procedure then enters an iterative sampling phase, which systematically executes intra-replica updates via Gibbs sampling (Algorithm S1), inter-replica configuration exchanges (Algorithm S2), adaptive temperature adjustments (Algorithm S3), and sample collection. Finally, all collected raw token sequences, denoted as $\mathcal{S}_{\text{raw}}$, are decoded into 3D atomic coordinates and assessed for structural confidence (Algorithm S4).

4.1.1 Initialization

Replica initialization in Parallel Tempering (PT) entails two key components: defining the initial structural states, denoted as $\mathbf{S}$, and establishing the temperature schedule, denoted as $\mathbf{T}$.

To initialize the structural state s_i for each of the N_{rep} replicas, the ESM3 model processes the input protein sequence **Seq** to generate the logits $\mathbf{z}$ for the structure tokens across all positions. Applying a softmax function to these logits yields a positional probability distribution. From this distribution, we independently sample initial conformations for all N_{rep} replicas, ensuring a diverse set of starting points for the conformational search.

Algorithm 1 Adaptive Parallel Tempering with Blocked Exchanges

Input:
Seq: Input amino acid sequence
Output:
$\mathcal{S}, \mathcal{X}, \mathcal{C}$: Sets of Tokens, Coordinates, and Confidence scores.

▷ Step 1: Initialization
$(\mathbf{T}, \mathbf{\Delta}, \mathbf{S}) \leftarrow \text{InitializeReplicas}(\mathbf{Seq})$ ▷ Initialize temperatures and states
$L \leftarrow \text{len}(\mathbf{Seq})$ ▷ Get sequence length
$\mathbf{B} \leftarrow \text{PartitionBlocks}(\mathbf{S}, N_{\text{nn}})$ ▷ Compute neighbor lists for block partitioning
$\mathcal{S} \leftarrow \emptyset$ ▷ Initialize set for token samples
▷ Step 2: Sampling Loop
for all $t \in \{1, \ldots, N_{\text{iter}}\}$ **do**
▷ (a) Update Nearest Neighbors (Periodically)
  if $t \pmod{N_{\text{nn_reset}}} = 0$ **then**
    $\mathbf{B} \leftarrow \text{PartitionBlocks}(\mathbf{S}, N_{\text{nn}})$
  end if
▷ (b) Intra-Replica Update (Parallel Gibbs)
  $\mathbf{S} \leftarrow \text{IntraReplicaUpdate}(\mathbf{S}, \mathbf{T}, \mathbf{B}, L)$ ▷ See Algorithm S1
▷ (c) Inter-Replica Exchange (Block exchange)
  $(\mathbf{S}, \boldsymbol{\alpha}_{\text{exchange}}) \leftarrow \text{InterReplicaExchange}(\mathbf{S}, \mathbf{T}, \mathbf{B}, t, L)$ ▷ See Algorithm S2
▷ (d) Update Observed Rates and Adapt Temperatures
  if $t \pmod 2 = 1$ **then** ▷ Odd step: update rates for odd pairs
    $\boldsymbol{\alpha}_{\text{curr}}[1:2:N_{\text{rep}}-1] \leftarrow \boldsymbol{\alpha}_{\text{exchange}}[1:2:N_{\text{rep}}-1]$
  else ▷ Even step: update rates for even pairs and then update T
    $\boldsymbol{\alpha}_{\text{curr}}[2:2:N_{\text{rep}}-1] \leftarrow \boldsymbol{\alpha}_{\text{exchange}}[2:2:N_{\text{rep}}-1]$
    $(\mathbf{T}, \mathbf{\Delta}) \leftarrow \text{UpdateTemperatures}(\mathbf{T}, \mathbf{\Delta}, \boldsymbol{\alpha}_{\text{curr}}, t)$ ▷ See Algorithm S3
  end if
▷ (e) Collect Sample
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathbf{S}[i] \mid i = 1..N_{\text{rep}}\}$ ▷ Add samples from all replicas
end for
▷ Step 3: Decode Structures
$\mathcal{X}, \mathcal{C} \leftarrow \emptyset, \emptyset$ ▷ Initialize output sets
for each $\mathbf{s}$ in $\mathcal{S}$ **do** ▷ Iterate through collected token samples
  $\mathbf{X} \leftarrow \text{DecodeTokensToCoords}(\mathbf{s})$ ▷ Decode tokens to 3D coordinates
  $c \leftarrow \text{CalculateConfidence}(\mathbf{s}, \mathbf{Seq})$ ▷ See Algorithm S4
  $\mathcal{X} \leftarrow \mathcal{X} \cup \{\mathbf{X}\}; \mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$ ▷ Add tuple to output sets
end for
return $\mathcal{S}, \mathcal{X}, \mathcal{C}$

Concurrently, we establish a temperature ladder, defined as $T_1 > T_2 > \cdots > T_{N_{\text{rep}}}$, across the N_{rep} replicas. This temperature schedule fundamentally dictates PT efficiency: the minimum temperature, $T_{\min}$, largely governs sampling accuracy, whereas the maximum temperature, $T_{\max}$, must be sufficiently elevated to cross high-energy barriers. Furthermore, the energy distributions of adjacent replicas (T_j, T_{j+1}) must exhibit adequate overlap to guarantee an appropriate exchange acceptance rate.

To increase sampling density in the low-temperature regime—a crucial requirement for navigating complex energy surfaces—we employ a non-uniform temperature schedule. Specifically, we introduce a concentration parameter $c = 2.0$ (where $c > 1$ skews the distribution toward lower temperatures) to dictate the N_{rep} temperature states. The procedure first defines a normalized spacing, u_i, as follows:

$$u_i = \left(\frac{i}{N_{\text{rep}}}\right)^c, \quad \text{for } i \in \{1, 2, \ldots, N_{\text{rep}}\} \tag{1}$$

Subsequently, these normalized intervals are mapped onto the logarithmic temperature domain, $[\log T_{\min}, \log T_{\max}]$. The final temperature T_i for replica i (ensuring $T_1 = T_{\max}$ and $T_{N_{\text{rep}}} = T_{\min}$) is given by:

$$T_i = \exp\left(\log T_{\min} + u_{N_{\text{rep}}-i+1} \cdot (\log T_{\max} - \log T_{\min})\right) \tag{2}$$

4.1.2 Intra-Replica Update

Within each replica i, intra-replica updates proceed via a block Gibbs sampling strategy (Algorithm S1). During a single Gibbs sweep, the algorithm iterates through all L residues according to a random permutation $\mathcal{O}$, ensuring every residue is targeted exactly once. For a given target residue j, we define a local sampling block, denoted as $b = \mathbf{B}[j]$. Inspired by the ESM3 encoder architecture—which relies on a restricted local structural neighborhood—this block comprises the target residue j alongside its $N_{\text{nn}} - 1$ spatially nearest neighbors. To establish this spatial proximity, the procedure decodes the structure tokens from the preceding iteration, computes the pairwise C_α-C_α distances, and identifies the spatially closest residues.

During the update step, the structure tokens for all residues in this block b are masked (Alg. S1). The protein sequence **Seq** and the remaining unmasked structure tokens s_i are provided as input to the ESM3, which then predicts the logits $\mathbf{z}$ for the masked positions. We apply a temperature-scaled softmax to these logits, using the replica's specific temperature T_i, and then sample the new structure tokens for the entire block b from the predicted categorical distribution:

$$s_i[b] \leftarrow \text{Sample}(\text{Softmax}(\mathbf{z}[b]/T_i)) \tag{3}$$

We chose this block-based approach over single-site updates to enhance sampling efficiency. Conformational differences often manifest as variations in local structures. By updating an entire block simultaneously, our method facilitates these concerted local transitions, thereby significantly improving the sampler's ability to explore the conformational space.

During the update phase, the structure tokens corresponding to all residues within the target block b are masked (Algorithm S1, Line 5). The protein sequence, **Seq**, along with the remaining unmasked structure tokens, s_i, are then passed to the ESM3 model to predict the logits $\mathbf{z}$ for the masked positions. Subsequently, we apply a temperature-scaled softmax function to these logits using the replica's designated temperature, T_i. New structure tokens for the entire block b are then sampled from

the resulting categorical distribution (Algorithm S1, Line 7):

$$s_i[b] \leftarrow \text{Sample}(\text{Softmax}(\mathbf{z}[b]/T_i)) \tag{4}$$

We deliberately favor this block Gibbs strategy over single-site updates to accelerate sampling convergence. Because distinct conformational states typically emerge from concerted variations in local structure, updating an entire spatial neighborhood simultaneously empowers the sampler to capture these coupled transitions efficiently. Consequently, this approach dramatically enhances the exploration of the overall conformational space.

4.1.3 Inter-Replica Exchange

During the inter-replica exchange phase (Algorithm S2), we propose local, partial configuration swaps between adjacent temperature replicas rather than exchanging their full, global states. This localized strategy addresses a fundamental constraint in statistical mechanics: the probability of accepting an exchange depends on the energy difference between replicas, a quantity that scales extensively with the system's degrees of freedom. Consequently, for high-dimensional systems such as proteins, attempting full-chain exchanges typically generates massive energy discrepancies, ultimately resulting in prohibitively low acceptance rates.

Our localized strategy mitigates this bottleneck, significantly enhancing the overall acceptance probability. Specifically, we randomly sample a subset of residue indices, I_{sel}, of size $N_{\text{exchange_blocks}}$. The effective exchange region, I_{exchange}, is subsequently defined as the union of the local spatial blocks corresponding to these selected indices:

$$I_{\text{exchange}} = \bigcup_{j \in I_{\text{sel}}} \mathbf{B}[j]$$

Consequently, when proposing a swap between two adjacent replicas, i and j, the algorithm exchanges structure tokens exclusively within this restricted domain, I_{exchange}.

The probability of accepting a proposed exchange, denoted as A, is governed by the Metropolis-Hastings criterion. This criterion evaluates the ratio between the probability of the proposed swapped state, P_{proposed}, and that of the current joint state, P_{current}. Let s_i and s_j represent the structure tokens within the designated exchange region, I_{exchange}, for adjacent replicas i and j, respectively. Because the replicas operate independently, the probability of the current joint configuration is the product of their individual probabilities:

$$P_{\text{current}} = P(s_i \mid \mathbf{p}_i)P(s_j \mid \mathbf{p}_j)$$

Similarly, the probability of the proposed configuration—where the tokens exclusively within I_{exchange} have been swapped—is given by:

$$P_{\text{proposed}} = P(s_j \mid \mathbf{p}_i)P(s_i \mid \mathbf{p}_j)$$

The final acceptance probability is therefore formulated as:

$$A = \min\left(1, \frac{P_{\text{proposed}}}{P_{\text{current}}}\right)$$

To ensure numerical stability during implementation, this calculation is executed in the logarithmic domain. Specifically, a LogLikelihood$(s, \mathbf{p}_k)$ function computes $\log P(s \mid \mathbf{p}_k)$, allowing the acceptance ratio to be safely evaluated via log-subtraction prior to exponentiation.

4.1.4 Adaptive Temperature Update

The efficiency of this exchange process relies critically on the configuration of the temperature ladder, $\mathbf{T}$. The optimal ladder is dictated by the protein's specific energy surface—a topological feature that is inherently unknown *a priori* and highly target-specific. Consequently, relying on a static temperature assignment is inevitably suboptimal. To resolve this limitation, we implement an adaptive tuning scheme (Algorithm S3) [44] that dynamically adjusts the logarithmic temperature intervals, $\mathbf{\Delta}$, to maintain a precise target exchange acceptance rate, α_{target} [45].

Specifically, the algorithm updates the logarithmic temperature intervals, $\mathbf{\Delta}$, by evaluating the discrepancy between the observed empirical acceptance rate, $\boldsymbol{\alpha}_{\text{curr}}[i]$, and the predefined target rate, α_{target}:

$$\mathbf{\Delta}[i] \leftarrow \mathbf{\Delta}[i] + \eta \cdot (\boldsymbol{\alpha}_{\text{curr}}[i] - \alpha_{\text{target}}) \tag{5}$$

Here, η represents a decaying step size defined as $\eta = \gamma_c'/(t+1)^{\gamma_\xi}$.

However, the original adaptive temperature algorithm [44] is susceptible to a phenomenon we term high-temperature drift. This pathology arises because the baseline method strictly enforces the target acceptance rate across all temperature scales, including the extreme high-temperature regime. Consequently, the procedure progressively widens the upper temperature intervals, driving the maximum temperature, $T_{\max}$, to diverge toward physically irrelevant values. Ultimately, this unbounded expansion introduces severe computational inefficiency without yielding any meaningful exploration benefits.

To counteract this divergence, we introduce two stabilizing mechanisms into the temperature update scheme (Algorithm S3). First, we enforce a hard upper bound on the logarithmic intervals, denoted as $\Delta_{\max}$. This strict constraint explicitly halts unbounded interval expansion, a measure critical for anchoring the extreme high-temperature regime of the ladder. Second, we implement an asymmetric step size via a dynamically scaled γ_c' parameter. This modification deliberately amplifies the corrective penalty for low acceptance rates (i.e., when $\boldsymbol{\alpha}_{\text{curr}}[i] < \alpha_{\text{target}}$), thereby accelerating the contraction of excessively large temperature gaps.

4.1.5 Sample Collection and Decoding

Finally, upon completing N_{iter} iterations, the structure tokens $\mathbf{s}$ sampled across all N_{rep} replicas are aggregated to form the raw ensemble, denoted as $\mathcal{S}_{\text{raw}}$. These

sequences then enter the final decoding phase. While the foundational ESM3 publication reported all-atom decoding capabilities, the currently available open-source model weights (`esm3-sm-open-v1`) restrict decoding strictly to backbone coordinates. Consequently, after translating the raw tokens into backbone structures $\mathbf{X}$, we utilize FlowPacker [46] for rigorous side-chain packing, thereby generating the final, fully atomistic conformational ensemble.

4.2 A Novel Confidence Metric based on Inverse Folding

Anfinsen's principle establishes that a protein's native three-dimensional structure is uniquely determined by its amino acid sequence [47]. Building upon this fundamental tenet, we hypothesize that any physically realistic conformation—including transient intermediates and metastable states—must exhibit high compatibility with the native sequence. Conversely, structurally aberrant or physically unrealistic conformations will inevitably introduce geometric features that violate the sequence-encoded physicochemical constraints.

Guided by this theoretical rationale, we propose that the conditional probability of the native sequence, **Seq**, given a sampled structure token sequence, $\mathbf{s}$, serves as an ideal, reference-free metric for structure-sequence self-consistency. Specifically, a high likelihood, $P(\mathbf{Seq} \mid \mathbf{s})$, indicates a plausible, physically realistic conformation. Conversely, a low likelihood implies a physically unviable structure, as its distorted backbone geometry would preferentially encode an amino acid sequence distinct from the native target. Accordingly, we formulate this sequence likelihood as our primary structural confidence metric, denoted as c:

$$c = P(\mathbf{Seq} \mid \mathbf{s})$$

To compute this metric (Algorithm S4), we leverage the inverse folding capabilities of the multimodal ESM3 model. For each conformation sampled in Section 4.1, we input its constituent structure tokens, $\mathbf{s}$, into the model to extract the conditional probability distribution, $P(\text{Sequence} \mid \mathbf{s})$. Subsequently, we compute the total log-likelihood of the target native sequence:

$$\mathcal{L}_{\text{total}} = \log P(\mathbf{Seq} \mid \mathbf{s})$$

To guarantee that this metric remains length-independent—and is therefore directly comparable across diverse protein targets—we normalize this value by the sequence length, L. This operation yields the final mean log-likelihood confidence score, c:

$$c = \frac{\mathcal{L}_{\text{total}}}{L}$$

4.3 Guided Sampling Strategy Integrating Prior Structural Information

The Parallel Tempering (PT) framework detailed in Section 4.1 is inherently designed for unbiased, global exploration of the full conformational energy surface. However,

in numerous practical applications, a reference conformation of the target protein is already established, whether through empirical determination or high-confidence structural prediction. Under these circumstances, the scientific objective shifts: rather than redundantly sampling this characterized state, the goal becomes the targeted discovery of alternative, previously undocumented conformations. To facilitate this, we introduce a guided sampling strategy that actively alters the ESM3 energy surface to strictly penalize exploration in the vicinity of the known state. Consequently, the algorithm pivots from sampling the original conditional distribution, $P(\text{Structure} \mid \text{Sequence})$, to sampling a modified, biased distribution, $P_{\text{guided}}(\text{Structure} \mid \mathbf{Seq}, \mathbf{s}_{\text{ref}})$, which is explicitly constructed to fill the corresponding energy well.

To operationalize this targeted exploration (Algorithm S5), we investigate two distinct logit suppression strategies. Both protocols necessitate initially encoding the known reference conformation into its discrete structure token sequence, defined as $\mathbf{s}_{\text{ref}} = \{s_{\text{ref},1}, \ldots, s_{\text{ref},L}\}$.

The first approach, denoted as Strategy 1 (Token-Level Suppression), introduces a minimal algorithmic intervention. During the Gibbs update of a spatial block containing sequence position p, the ESM3 model generates a predictive logit tensor, $\mathbf{z}$. Under this strategy, we explicitly mask $\mathbf{z}$ by assigning a value of $-\infty$ to the specific logit associated with the reference token, $s_{\text{ref},p}$.

The second approach, Strategy 2 (Cluster-Level Suppression), penalizes not only the specific reference token but also a broader neighborhood of structurally analogous tokens. Implementing this wider suppression necessitates a prerequisite offline clustering phase to group tokens based on their latent embedding representations. Specifically, we extract the continuous token embeddings from the ESM3 architecture, project them into a lower-dimensional space via UMAP [36], and subsequently partition them using HDBSCAN density-based clustering [48, 49]. To optimize this partitioning, we execute a comprehensive hyperparameter grid search (Table 6). The generated clustering configurations are then stratified into ten discrete bins according to their mean cluster size. Within each stratum, we isolate the configuration that maximizes the Density-Based Clustering Validation (DBCV) score. Ultimately, this stratified selection yields a diverse ensemble of the ten optimal clustering solutions, denoted as $\mathcal{M}$ (Table 7).

Under Strategy 2, during the Gibbs update of sequence position p, the algorithm identifies the reference token, $s_{\text{ref},p}$, alongside its corresponding pre-computed cluster, $\mathcal{S}_{\text{cluster}}$. Crucially, we explicitly mask the logits for both $s_{\text{ref},p}$ and all other structurally analogous tokens $s \in \mathcal{S}_{\text{cluster}}$ by assigning them a value of $-\infty$ (Algorithm S5, Lines 12–15). By driving the sampling probability of the entire reference local structure to zero, this position-specific, cluster-based masking operation robustly suppresses the $\mathbf{s}_{\text{ref}}$ signal. Consequently, it forces the sampler to navigate toward alternative, distinct conformational basins.

Table 6: Grid Search Parameters for UMAP Dimensionality Reduction and HDBSCAN Clustering

Algorithm	Parameter	Candidate Values
UMAP	`n_neighbors`	[2, 3, 5, 8]
	`min_dist`	[0.0]
	`n_components`	[128, 256]
	`metric`	[cosine]
HDBSCAN	`min_cluster_size`	[2, 3, 4, 5]
	`min_samples`	[2, 3]
	`cluster_selection_epsilon`	[0.05, 0.1, 0.2]
	`metric`	[euclidean]

Table 7: Clustering Results (Top 10 by DBCV)

dbcv	n_clusters	n_noise	noise_ratio	avg_cluster_size
0.810	790	544	0.133	4.496
0.703	100	52	0.013	40.440
0.630	612	415	0.101	6.015
0.627	542	288	0.070	7.026
0.626	408	374	0.091	9.123
0.604	237	107	0.026	16.831
0.603	530	1214	0.296	5.438
0.601	337	337	0.082	11.154
0.590	126	323	0.079	29.944
0.574	241	489	0.119	14.967

Declarations

- **Data availability**
  Data will be available upon publication.
- **Code availability**
  Code will be available upon publication.
- **Competing interests**
  The authors declare no competing interests.
- **Funding**
  This work was supported by the National Key Research and Development Program of China (2024YFC3405500) and the National Natural Science Foundation of China (32271297, 82130055).

References

[1] Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).

[2] Wankowicz, S. A. *et al.* Automated multiconformer model building for x-ray crystallography and cryo-em. *Elife* **12**, RP90606 (2024).

[3] Meisburger, S. P., Thomas, W. C., Watkins, M. B. & Ando, N. X-ray scattering studies of protein structural dynamics. *Chemical reviews* **117**, 7615–7672 (2017).

[4] Grandori, R. Protein structure and dynamics in the era of integrative structural biology. *Frontiers in Biophysics* **1**, 1219843 (2023).

[5] Ellaway, J. I. *et al.* Identifying protein conformational states in the protein data bank: toward unlocking the potential of integrative dynamics studies. *Structural Dynamics* **11** (2024).

[6] Nwanochie, E. & Uversky, V. N. Structure determination by single-particle cryo-electron microscopy: only the sky (and intrinsic disorder) is the limit. *International journal of molecular sciences* **20**, 4186 (2019).

[7] Laurents, D. V. Alphafold 2 and nmr spectroscopy: partners to understand protein structure, dynamics and function. *Frontiers in molecular biosciences* **9**, 906437 (2022).

[8] Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).

[9] Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**, 1129–1143 (2018).

[10] Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *nature* **596**, 583–589 (2021).

[11] Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

[12] Evans, R. *et al.* Protein complex prediction with alphafold-multimer. *biorxiv* 2021–10 (2021).

[13] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024).

[14] Krishna, R. *et al.* Generalized biomolecular modeling and design with rosettafold all-atom. *Science* **384**, eadl2528 (2024).

[15] Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with alphafold2. *Elife* **11**, e75751 (2022).

[16] Chakravarty, D. & Porter, L. L. Alphafold2 fails to predict protein fold switching. *Protein Science* **31**, e4353 (2022).

[17] Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and alphafold2. *Nature* **625**, 832–839 (2024).

[18] Song, Y. *et al.* Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

[19] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).

[20] Jing, B. *et al.* Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198* (2023).

[21] Lu, J., Zhong, B., Zhang, Z. & Tang, J. Str2str: A score-based framework for zero-shot protein conformation sampling. *arXiv preprint arXiv:2306.03117* (2023).

[22] Jing, B., Berger, B. & Jaakkola, T. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845* (2024).

[23] Lu, J. *et al.* Structure language models for protein conformation generation. *arXiv preprint arXiv:2410.18403* (2024).

[24] Chakravarty, D. *et al.* Alphafold predictions of fold-switched conformations are driven by structure memorization. *Nature communications* **15**, 7296 (2024).

[25] Kalakoti, Y. & Wallner, B. Afsample2 predicts multiple conformations and ensembles with alphafold2. *Communications Biology* **8**, 373 (2025).

[26] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

[27] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

[28] Su, J. *et al.* Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv* 2023–10 (2023).

[29] Van Kempen, M. *et al.* Fast and accurate protein structure search with foldseek. *Nature biotechnology* **42**, 243–246 (2024).

[30] Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* eads0018 (2025).

[31] Van Den Oord, A., Vinyals, O. *et al.* Neural discrete representation learning. *Advances in neural information processing systems* **30** (2017).

[32] Swendsen, R. H. & Wang, J.-S. Replica monte carlo simulation of spin-glasses. *Physical review letters* **57**, 2607 (1986).

[33] Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters* **314**, 141–151 (1999).

[34] Saldaño, T. *et al.* Impact of protein conformational diversity on alphafold predictions. *Bioinformatics* **38**, 2742–2748 (2022).

[35] Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**, 702–710 (2004).

[36] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[37] Wang, Y. *et al.* Protein conformation generation via force-guided se (3) diffusion models. *arXiv preprint arXiv:2403.14088* (2024).

[38] Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026–1028 (2017).

[39] Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature methods* **19**, 1109–1115 (2022).

[40] Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* **22**, 2577–2637 (1983).

[41] Joosten, R. P. *et al.* A series of pdb related databases for everyday needs. *Nucleic acids research* **39**, D411–D419 (2010).

[42] Liu, Z., Feng, B., Cao, H. & Li, Y. From static structures to ensembles: Studying and harnessing protein structure tokenization. *arXiv preprint arXiv:2511.10056* (2025).

[43] Yuan, X., Wang, Z., Collins, M. & Rangwala, H. Protein structure tokenization: Benchmarking and new recipe. *arXiv preprint arXiv:2503.00089* (2025).

[44] Miasojedow, B., Moulines, E. & Vihola, M. An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics* **22**, 649–664 (2013).

[45] Atchadé, Y. F., Roberts, G. O. & Rosenthal, J. S. Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Statistics and Computing* **21**, 555–568 (2011).

[46] Lee, J. S. & Kim, P. M. Flowpacker: protein side-chain packing with torsional flow matching. *Bioinformatics* **41**, btaf010 (2025).

[47] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).

[48] Campello, R. J., Moulavi, D. & Sander, J. *Density-based clustering based on hierarchical density estimates.* Pacific-Asia conference on knowledge discovery and data mining, 160–172 (Springer, 2013).

[49] McInnes, L. & Healy, J. *Accelerated hierarchical density based clustering.* 2017 IEEE international conference on data mining workshops (ICDMW), 33–42 (IEEE, 2017).

Supplementary Material

Appendix A Supplementary Tables

Table S1: Notation Table

Symbol	Description
Core PT Variables	
N_{rep}	Total number of replicas.
N_{iter}	Total number of MCMC iterations.
t	Current MCMC iteration index.
$\mathbf{T}$	Vector of temperatures ($T_1 > \cdots > T_{N_{\text{rep}}}$).
T_i	Temperature of replica i.
T_{min}	Minimum (coldest) temperature ($T_{N_{\text{rep}}}$).
$\mathbf{\Delta}$	Vector of log-intervals between adjacent temperatures.
α_{target}	Target exchange acceptance rate for temperature adaptation.
Protein & State Variables	
Seq	Input amino acid sequence.
L	Sequence length ($L = \text{len}(\mathbf{Seq})$).
$\mathbf{S}$	State matrix ($N_{\text{rep}} \times L$) of structure tokens for all replicas.
s_i	The state (token sequence) of replica i (i.e., $\mathbf{S}[i]$).
$\mathbf{s}$	A single structure token sequence (a sample).
$\mathbf{X}$	A set of 3D coordinates corresponding to a sample $\mathbf{s}$.
c	Confidence score for a sample $\mathbf{s}$.
MCMC Sampling Variables	
$\mathbf{B}$	Block partition map; $\mathbf{B}[j]$ is the neighbor list for residue j.
N_{nn}	Number of nearest neighbors defining a block.
b	A single block (set of token indices) for intra-replica updates.
$N_{\text{exchange_blocks}}$	Number of blocks used in a joint inter-replica exchange.
I_{exchange}	Set of all token indices involved in an inter-replica exchange.
$\mathbf{z}$	Logits predicted by ESM3.
$\boldsymbol{\alpha}_{\text{exchange}}, \boldsymbol{\alpha}_{\text{out}}$	Vector of acceptance rates from the most recent exchange attempt.
$\boldsymbol{\alpha}_{\text{curr}}$	Stored vector of acceptance rates used for adaptation.
A	Acceptance probability for a single replica pair exchange.
Guided Sampling Variables	
$\mathbf{s}_{\text{ref}}$	Reference structure token sequence to be suppressed.
$\mathcal{M}$	Pre-computed cluster map for structure tokens.
$\mathcal{S}_{\text{cluster}}$	Set of tokens belonging to a specific cluster in $\mathcal{M}$.
Confidence & Likelihood Variables	
$\mathcal{L}_{\text{total}}$	Total log-likelihood of **Seq** given $\mathbf{s}$.
$\mathcal{L}_{\text{avg}}$	Average per-residue log-likelihood (confidence score c).

Continued on next page

Table S1 – continued from previous page

Symbol	Description
Output Set Variables	
$\mathcal{S}_{\text{raw}}$	Set of raw token samples collected from all replicas.
$\mathcal{S}, \mathcal{X}, \mathcal{C}$	Output sets of decoded samples, coordinates, and confidence scores.

Appendix B Supplementary Algorithms

Algorithm S1 Intra-Replica Update

def IntraReplicaUpdate($\mathbf{S}, \mathbf{T}, \mathbf{B}, L$):

for all replicas $i \in \{1, \ldots, N_{\text{rep}}\}$ **in parallel do**
  $\mathcal{O} \leftarrow$ RandomPermutation($\{1, \ldots, L\}$) ▷ Random update order of residues
  for each residue index j in $\mathcal{O}$ **do**
    $b \leftarrow \mathbf{B}[j]$ ▷ Block b is residue j and its $N_{\text{nn}} - 1$ neighbors
    $s_i[b] \leftarrow$ MASK ▷ Mask all tokens in the block
    $\mathbf{z} \leftarrow$ StructureTokenLogits(s_i, b) ▷ Get logits for state s_i with block b masked
    $s_i[b] \leftarrow$ Sample(Softmax($\mathbf{z}[b]/T_i$)) ▷ Sample new tokens for block b
  end for
end for
return S

Algorithm S2 Inter-Replica Exchange

def InterReplicaExchange($\mathbf{S}, \mathbf{T}, \mathbf{B}, t, L, N_{\text{exchange_blocks}} = 3$) :

▷ Define the joint token set I_{exchange} to be exchanged

$I_{\text{sel}} \leftarrow \textsc{SampleIndices}(\{1, \ldots, L\}, N_{\text{exchange_blocks}})$ ▷ 1. Sample $N_{\text{exchange_blocks}}$ residue indices

$I_{\text{exchange}} \leftarrow \bigcup_{j \in I_{\text{sel}}} \mathbf{B}[j]$ ▷ 2. Get token indices from all blocks defined by I_{sel}

▷ Calculate logits for the masked block

$\mathbf{S}' \leftarrow \mathbf{S}$ with $[:, I_{\text{exchange}}]$ set to MASK ▷ Define a temporarily masked state

$\mathbf{z} \leftarrow \textsc{StructureTokenLogits}(\mathbf{S}', I_{\text{exchange}})$ ▷ Get logits using the masked state

▷ Define exchange pairs based on odd/even step t

if t is even **then**

  $I_t \leftarrow \{1, 3, 5, \ldots\}$ ▷ Set pairs (1,2), (3,4), ...

else

  $I_t \leftarrow \{2, 4, 6, \ldots\}$ ▷ Set pairs (2,3), (4,5), ...

end if

$\boldsymbol{\alpha}_{\text{out}} \leftarrow \text{zeros}(N_{\text{rep}} - 1)$ ▷ Initialize acceptance rate vector

for all replicas $i \in I_t$ **in parallel do**

  $j \leftarrow i + 1$

▷ Calculate log-likelihood for current and proposed states

  $\mathbf{p}_i \leftarrow \text{Softmax}\,(\mathbf{z}[i]/T_i)$ ▷ Calculate probability distribution for i

  $\mathbf{p}_j \leftarrow \text{Softmax}\,(\mathbf{z}[j]/T_j)$ ▷ Calculate probability distribution for j

  $\log P_{\text{current}} \leftarrow \textsc{LogLikelihood}(s_i[I_{\text{exchange}}], \mathbf{p}_i) + \textsc{LogLikelihood}(s_j[I_{\text{exchange}}], \mathbf{p}_j)$

  $\log P_{\text{proposed}} \leftarrow \textsc{LogLikelihood}(s_j[I_{\text{exchange}}], \mathbf{p}_i) + \textsc{LogLikelihood}(s_i[I_{\text{exchange}}], \mathbf{p}_j)$

  $A \leftarrow \min(1, \exp(\log P_{\text{proposed}} - \log P_{\text{current}}))$ ▷ Metropolis-Hastings acceptance rate

  $\boldsymbol{\alpha}_{\text{out}}[i] \leftarrow A$ ▷ Store acceptance rate for interval i

  if $\textsc{Uniform}(0, 1) < A$ **then**

    $\textsc{exchange}(s_i[I_{\text{exchange}}], s_j[I_{\text{exchange}}])$ ▷ Accept exchange

  end if

end for

return $\mathbf{S}, \boldsymbol{\alpha}_{\text{out}}$

Algorithm S3 Adaptive Temperature Update

def UpdateTemperatures($\mathbf{T}, \mathbf{\Delta}, \boldsymbol{\alpha}_{\text{curr}}, t, \alpha_{\text{target}=0.234}, \gamma_c = 0.25, \gamma_\xi = 0.6, \Delta_{\max} = 2$)

for all intervals $i \in \{1, \ldots, N_{\text{rep}} - 1\}$ **in parallel do**
  diff $\leftarrow \boldsymbol{\alpha}_{\text{curr}}[i] - \alpha_{\text{target}}$ ▷ Calculate difference from target
  if diff < 0 **then**
    $\gamma_c' \leftarrow (1 - \alpha_{\text{target}})/\alpha_{\text{target}}$
  else
    $\gamma_c' \leftarrow \gamma_c$
  end if
  $\eta \leftarrow (\gamma_c'/(t+1)^{\gamma_\xi})$ ▷ Learning rate
  $\mathbf{\Delta}[i] \leftarrow \mathbf{\Delta}[i] + \eta \cdot \text{diff}$ ▷ Update log-intervals
end for
$\mathbf{\Delta} \leftarrow \textsc{Clamp}(\mathbf{\Delta}, \max = \Delta_{\max})$ ▷ Enforce maximum log-interval
▷ Reconstruct temperatures from intervals
$\mathbf{T}[N_{\text{rep}}] \leftarrow T_{\min}$
for $i \leftarrow N_{\text{rep}} - 1$ down to 1 **do**
  $\mathbf{T}[i] \leftarrow \mathbf{T}[i+1] + \exp(\mathbf{\Delta}[i])$
end for
return $(\mathbf{T}, \mathbf{\Delta})$

Algorithm S4 Calculate Confidence (via Average Log-Likelihood)

def CalculateConfidence($\mathbf{s}, \mathbf{Seq}$):
▷ Computes the per-residue log-likelihood of the true sequence given the structure
$\mathbf{z} \leftarrow \textsc{SequenceLogits}(\mathbf{s})$ ▷ Get sequence logits from structure $\mathbf{s}$
$\mathcal{L}_{\text{total}} \leftarrow \textsc{ComputeLogLikelihood}(\mathbf{z}, \mathbf{Seq})$ ▷ Sum of log-probs
$L \leftarrow \textsc{len}(\mathbf{Seq})$
$\mathcal{L}_{\text{avg}} \leftarrow \mathcal{L}_{\text{total}}/L$ ▷ Normalize by length
return $\mathcal{L}_{\text{avg}}$

Algorithm S5 Guided Sampling Logit Suppression Strategies

```
def ApplyLogitSuppression(s_ref, M, z, mode):
    ▷ s_ref: ref. token sequence, M: cluster map, z: logits, mode: 1 (Token) or 2 (Cluster)
L ← len(s_ref)                                  ▷ Get sequence length
for all p in 0 ... L − 1 do                     ▷ Iterate over all positions
    s_ref,p ← s_ref[p]                          ▷ Get the known reference token for position p
    if mode == 1 then
                                                ▷ Strategy 1: Token-Level Suppression
        z[p, s_ref,p] ← −∞                      ▷ Set logit for s_ref,p at pos p to -inf
    else if mode == 2 then
                                                ▷ Strategy 2: Cluster-Level Suppression
        S_cluster ← GetCluster(M, s_ref,p)      ▷ Get cluster from map M
        for all s in S_cluster do
            z[p, s] ← −∞                        ▷ Set logit for all tokens in cluster at pos p to -inf
        end for
    end if
end for
return z                                        ▷ Return modified logits for all positions
```

Appendix C Supplementary Figures

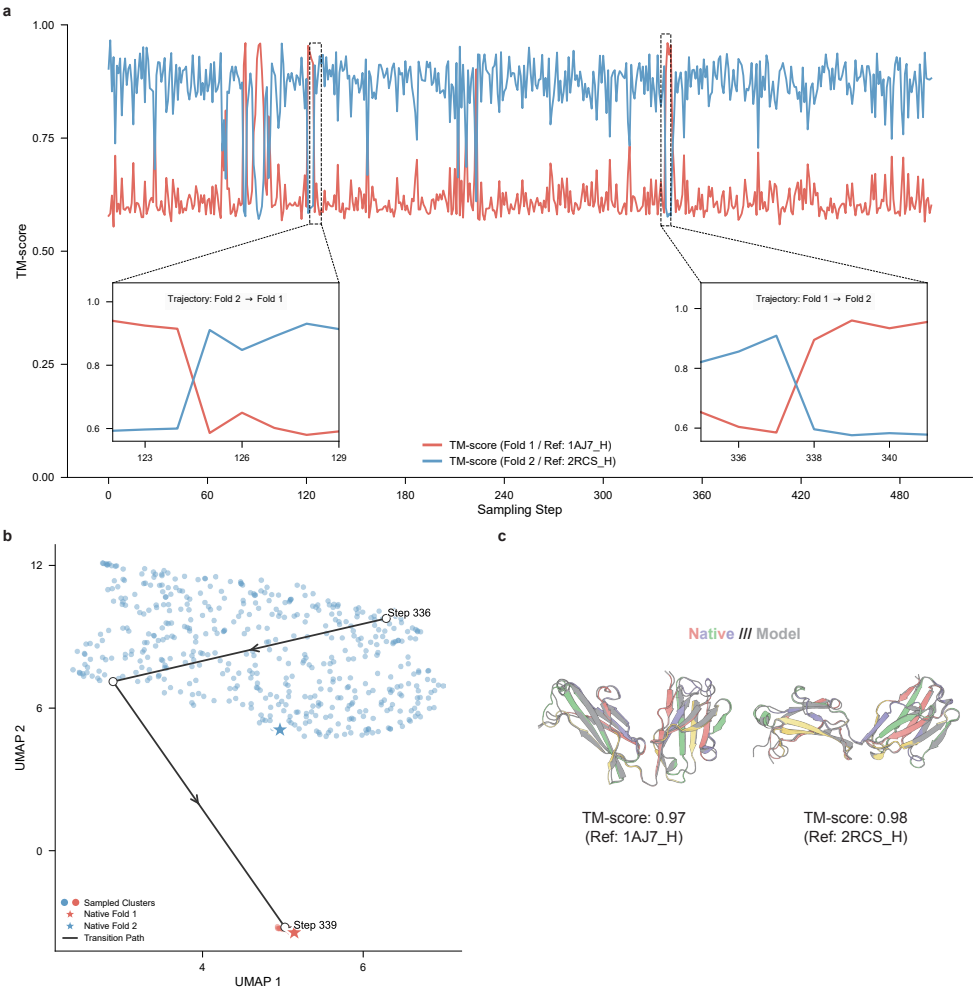

Fig. S1: **MSFold captures both native states of the Antibody 48G7 Fab. a**, Evolution of structural similarity (TM-score) relative to the germline state (Fold 1, PDB: 1AJ7_H; red) and the affinity-matured state (Fold 2, PDB: 2RCS_H; blue). The trajectory consistently samples high-accuracy structures of either conformation, rather than intermediate mixtures. Two insets highlight the reversible transitions between these states during the sampling process (e.g., Fold 2 → Fold 1 and Fold 1 → Fold 2). **b**, UMAP projection of the sampled conformational space. Each point represents a sampled structure, revealing two clusters corresponding to the native references (stars). The arrowed path tracks the specific transition event shown in the right inset of **a** (Steps 336–339). **c**, Structural superpositions of predicted models (grey) onto native references (colored). The predicted structures exhibit high structural similarity to the native states, achieving TM-scores of 0.97 and 0.98, respectively.

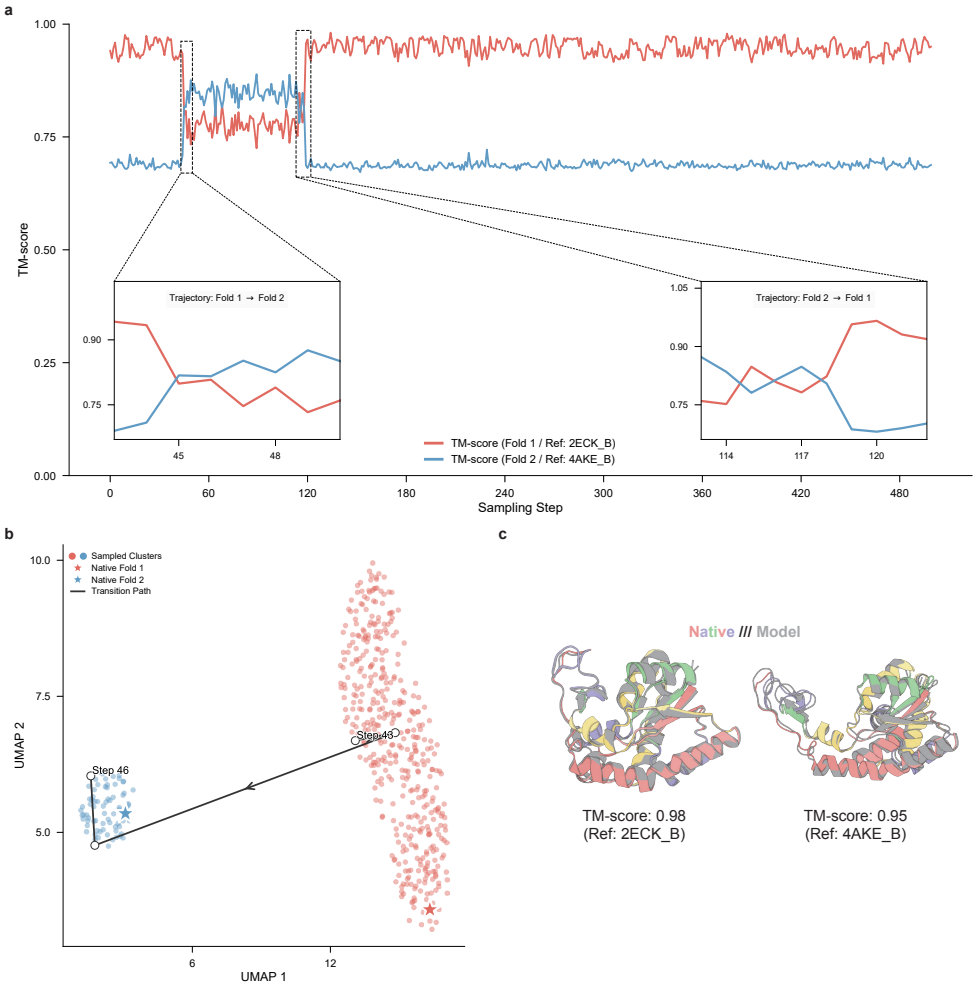

Fig. S2: **MSFold captures both native states of Adenylate Kinase. a**, Evolution of structural similarity (TM-score) relative to Fold 1 (PDB: 2ECK_B; red) and Fold 2 (PDB: 4AKE_B; blue). The trajectory demonstrates the model's ability to sample high-accuracy structures for both distinct conformations, effectively switching between them rather than lingering in intermediate states. Insets highlight the reversible transition events (e.g., Fold 1 → Fold 2) observed during the sampling run. **b**, UMAP projection of the sampled conformational space. The plot reveals two well-separated clusters corresponding to the native reference states (stars). The arrowed path visualizes a specific transition trajectory (Steps 43–46) connecting the two basins, corresponding to the event detailed in the left inset of **a**. **c**, Structural superpositions of generated models (grey) onto the native references (colored). MSFold reconstructs both conformations with high fidelity, achieving TM-scores of 0.98 and 0.95 for Fold 1 and Fold 2, respectively.