SF-Cluster: Frustration-Aware MSA Subsampling for Protein Conformational Control

Hanqun Cao [1 *] **Zijun Gao** [1 *] **Chunbin Gu** [1] **Ge Liu** [2] **Pheng Ann Heng** [1] **Pranam Chatterjee** [3 4]

Abstract

Protein structure predictors are sensitive to their multiple sequence alignment (MSA) input, making MSA subsampling a practical strategy for recovering alternative conformations. Existing approaches such as AF-Cluster operate in sequence space, supporting broad exploration but providing limited control over which conformational basin is sampled. We introduce **SF-Cluster**, a framework that uses frustration patterns to guide state-directed conformational sampling, together with a coverage-aware refinement step that prevents collapse toward dominant states. On fold-switching benchmarks, SF-Cluster improves targeted recovery of alternative conformations over sequence-space baselines, and successful subsets reflect protein-specific organization in frustration space. When the input MSA is structurally single-basin, no frustration-based strategy recovers non-reference conformations, indicating that subsampling can enrich conformations already represented within an MSA but cannot recover states absent from the underlying sequence pool. Our results establish frustration-pattern representations as a useful approach for controllable conformational sampling with existing structure prediction systems.

1. Introduction

Conformational heterogeneity is central to protein function. Fold-switching proteins, allosteric regulators, and receptor systems depend on transitions between structurally distinct states. Despite the success of AlphaFold2 (Jumper et al., 2021) in single-structure prediction, recovering alternative conformations from sequence alone remains difficult. Predictions are typically biased toward dominant states, and directing inference toward a specific conformational basin generally requires modifying the standard prediction pipeline (Chakravarty & Porter, 2022; Chakravarty et al., 2024).

MSA subsampling has emerged as a practical approach for increasing conformational diversity. Because AlphaFold2 is sensitive to its MSA input, changes to the MSA can alter the conformational basin sampled during prediction (Del Alamo et al., 2022; Stein & Mchaourab, 2022). AF-Cluster demonstrated this effect by partitioning an MSA according to sequence similarity and running structure prediction on each subset, showing that cluster-specific MSAs can recover multiple conformations for fold-switching proteins (Wayment-Steele et al., 2024). Subsequent methods have refined this strategy through sequence purification and stochastic inference (Xing et al., 2025; Bryant & Noé, 2024; Lee et al., 2025; Li et al., 2026). However, these approaches operate in sequence or perturbation space, where sequence similarity serves as an indirect surrogate for conformational state. As a result, they provide limited control over selecting MSA subsets associated with a desired structural state.

A more informative representation may be encoded in local energetic frustration. Frustration quantifies how well residue interactions are satisfied relative to alternative configurations and has been linked to conformational strain, switching behavior, and alternative fold preferences (Leusch et al., 2026). Frustration patterns across homologous sequences therefore provide information about conformational preference that is complementary to sequence similarity. We introduce **SF-Cluster**, an MSA subsampling framework that uses frustration-derived representations to guide state-directed conformational sampling. SF-Cluster also incorporates a coverage-aware refinement procedure that preserves low-frequency states that would otherwise be discarded by confidence-based ranking.

Across fold-switching benchmarks, SF-Cluster improves recovery of target conformations relative to AF-Cluster. These improvements are not explained by sequence diversity

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong [2]Department of Computer Science, University of Illinois at Urbana-Champaign [3]Department of Bioengineering, University of Pennsylvania [4]Department of Computer and Information Science, University of Pennsylvania. Correspondence to: Pranam Chatterjee <pranam@seas.upenn.edu>.

Published in the *AI for Science Workshop* (Oral) at the 43rd International Conference on Machine Learning (ICML 2026), Seoul, South Korea.

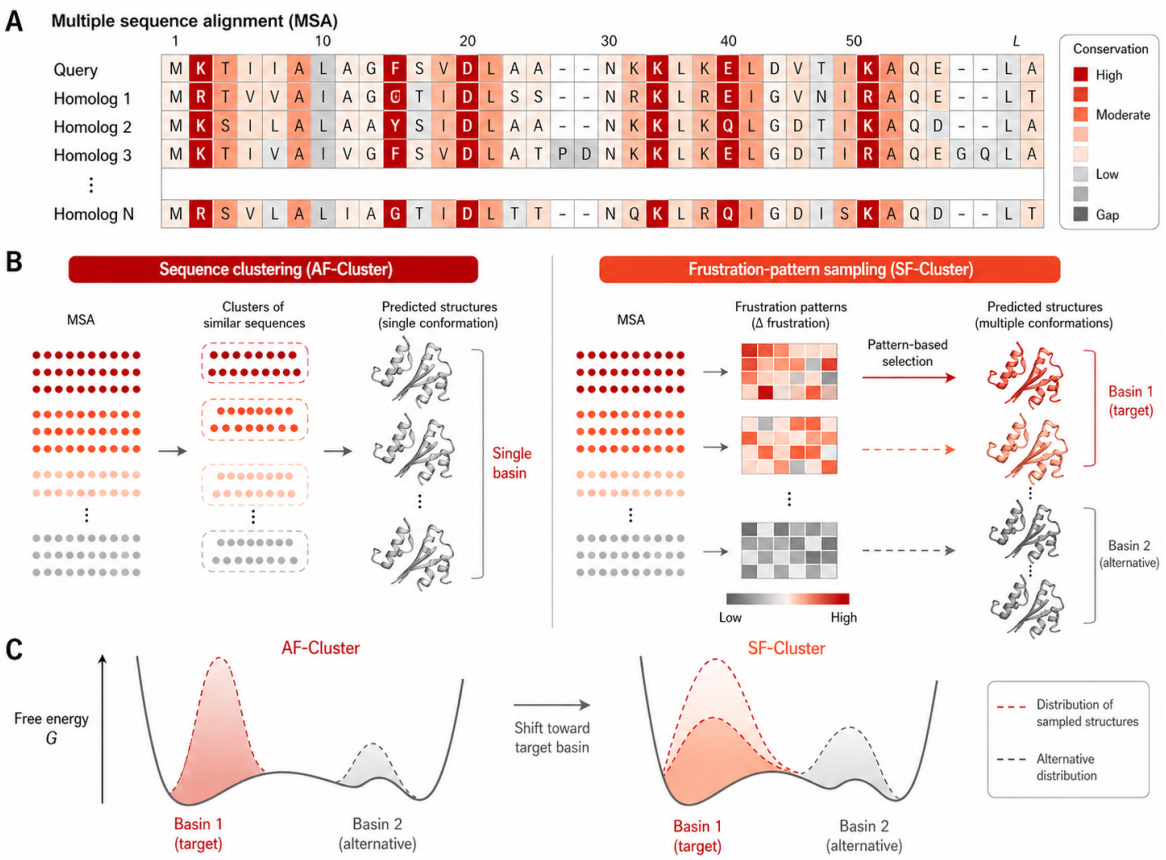

Figure 1. **Overview of SF-Cluster. (A)** A multiple sequence alignment encodes evolutionary variation across N homologs; residue conservation is shown by color intensity. **(B)** AF-Cluster partitions the MSA by sequence similarity, producing cluster-specific subsets that predominantly sample a single conformational basin. SF-Cluster instead computes per-homolog frustration patterns and selects MSA subsets according to their geometry in frustration-pattern space, directing predictions toward distinct conformational basins. **(C)** Viewed as basin reweighting on a free energy landscape, sequence-space clustering leaves the dominant basin largely unchanged, whereas frustration-pattern sampling shifts the predicted structural distribution toward the target basin while retaining coverage of the alternative.

alone. Successful MSA subsets are organized by protein-specific frustration patterns whose geometry differs across systems, indicating that conformational signals encoded within an MSA occupy preferred directions in frustration space. When the input MSA contains only a single structural basin, frustration-guided subsampling does not recover alternative conformations, defining a practical limit of what MSA subsampling can achieve.

Together, these results suggest a three-stage view of MSA subsampling: broad exploration in sequence space, state-directed focusing through frustration-pattern sampling, and retention of low-frequency conformations through coverage-aware refinement. Our contributions are:

- **Frustration-pattern representations for MSA subsampling.** We show that frustration-derived features provide a useful basis for MSA selection and encode conformational information that is not captured by sequence similarity alone.
- **Protein-specific geometry of conformational signals.** Across fold-switching benchmarks, successful subsampling strategies depend on the structure of frustration-pattern space within each protein family. Maximizing sequence diversity alone is insufficient for targeted conformational recovery.
- **Coverage-aware retention and limits of MSA subsampling.** We combine frustration-guided subsampling with coverage-aware refinement and show that MSA subsampling can enrich conformations already represented within an MSA, but cannot recover conformations absent from a structurally single-basin sequence pool.

2. Related Work

MSA perturbation for conformational sampling. The MSA plays a central role in determining which conformational basin AlphaFold2 samples during inference. Reduc-

ing MSA depth weakens dominant-state coevolutionary constraints and can increase recovery of alternative conformations (Del Alamo et al., 2022). AF-Cluster (Wayment-Steele et al., 2024) demonstrated that partitioning a natural MSA by sequence similarity can yield subsets associated with distinct structural states. Subsequent work has refined MSA selection through column mutagenesis (Stein & Mchaourab, 2022), sequence purification (Xing et al., 2025), and co-evolutionary disentanglement (Li et al., 2026). Other approaches condition prediction directly on state-associated sequence features (Bryant & Noé, 2024; Lee et al., 2025). At the same time, studies of AlphaFold2 failure modes have shown that fold-switch predictions are often driven by training-set memorization rather than inference over alternative conformational states (Chakravarty & Porter, 2022; Chakravarty et al., 2024). Despite these advances, existing MSA perturbation methods operate primarily in sequence space and provide limited guidance for selecting subsets associated with a desired conformational state.

Generative MSA augmentation. A related line of work generates or optimizes MSAs directly. EvoGen (Zhang et al., 2023), MSA-Generator (Zhang et al., 2024), MSAGPT (Chen et al., 2024), PLAME (Cao et al., 2025), and related methods (Sgarbossa et al., 2023; Venkatraman et al.) construct synthetic evolutionary contexts to improve downstream structure prediction. These studies demonstrate that modifying evolutionary context can substantially alter the structures explored by AlphaFold2. However, existing approaches do not explicitly connect variation within an MSA to the energetic preferences of homologous sequences across conformational states. SF-Cluster addresses this gap by using frustration-derived representations to identify homolog subsets with distinct conformational preferences.

3. Method

3.1. Problem Setup

Let $\mathcal{A} = \{\mathbf{s}_i \in \mathcal{V}^L\}_{i=1}^N$ denote a multiple sequence alignment of N homologs over an alphabet $\mathcal{V}$. A structure predictor f maps a subset $\mathcal{A}' \subseteq \mathcal{A}$ to a predicted structure $\mathbf{X} = f(\mathcal{A}', \xi)$, where ξ captures stochastic inference. We view f as inducing a distribution $P(\mathbf{X} \mid \mathcal{A}')$ over structure space, partitioned into conformational basins $\mathcal{B} = \{B_1, \ldots, B_S\}$ with basin probabilities

$$P(B_s \mid \mathcal{A}') = \int_{B_s} P(\mathbf{X} \mid \mathcal{A}')\, d\mathbf{X}. \tag{1}$$

The goal of targeted conformational focusing is to select $\mathcal{A}'$ such that $P(B_{s^*} \mid \mathcal{A}')$ is maximized for a desired basin B_{s^*} without exhaustive search over MSA subsets.

3.2. Representation-Induced Basin Reweighting

We assume that basin probabilities depend on a representation $\phi(\mathcal{A}')$ derived from MSA statistics:

$$P(B_s \mid \mathcal{A}') \propto \exp\big(g_s(\phi(\mathcal{A}'))\big), \tag{2}$$

where $g_s(\cdot)$ is an unknown basin-specific score function. Under this model, MSA subsampling changes $\phi(\mathcal{A}')$, which alters the relative probabilities assigned to conformational basins. The representation ϕ therefore determines the extent to which subsampling can bias prediction toward a desired state.

Methods such as AF-Cluster use sequence similarity as ϕ, partitioning an MSA into groups that are coherent in sequence space. We instead use a representation derived from local energetic frustration.

3.3. Why Frustration Patterns are State-Relevant

Let $Y \in \{1, \ldots, S\}$ denote the latent conformational basin associated with a homologous sequence $\mathbf{s}$. A representation is useful for subsampling when it contains information about Y. Sequence-based approaches implicitly assume that global sequence similarity captures this information:

$$P(Y \mid \mathbf{s}_i, \mathbf{s}_j) \approx P(Y \mid d_{\text{seq}}(\mathbf{s}_i, \mathbf{s}_j)). \tag{3}$$

This assumption can fail when a small number of mutations alter conformational energetics while preserving high sequence identity. GB98 T25I/L20A provides such an example: two substitutions eliminate the fold-switch signal while leaving overall sequence similarity to the wild-type sequence largely unchanged.

Motivated by this observation, we seek a representation $F_i \in \mathbb{R}^L$ satisfying

$$I(F_i; Y) \gg I(d_{\text{seq}}(\mathbf{s}_i, \cdot); Y), \tag{4}$$

where $I(\cdot;\cdot)$ denotes mutual information. We use the per-residue frustration index as F_i. Regions with elevated or redistributed frustration frequently coincide with hinges, interfaces, and switching regions whose energetic balance differs across conformational states. Similarity in frustration-pattern space therefore provides information about conformational preference that is not necessarily captured by sequence similarity. The region-level compression used in practice is defined in Section 3.5.

3.4. SF-Cluster: Frustration-Guided Subsampling

Local frustration quantifies how compatible a residue interaction is with the folded structure relative to alternative configurations. For each homolog $\mathbf{s}_i$, we compute a per-residue frustration vector $F_i \in \mathbb{R}^L$ using a predicted-structure-based frustration estimator, yielding the representation

$$\Phi(\mathcal{A}) = \{F_i\}_{i=1}^N. \tag{5}$$

Algorithm 1 SF-Cluster / Pattern-SF

Require: MSA $\mathcal{A} = \{\mathbf{s}_i\}_{i=1}^N$, structure predictor f, frustration predictor h, region partition $\mathcal{R} = \{R_k\}_{k=1}^K$, number of subsets M, samples per subset K, per-basin budget K_b
Ensure: Refined structure ensemble $\mathcal{D}^*$

```
// Step 1: compute frustration representations
for i = 1 to N do
    F_i ← h(s_i) {per-residue frustration profile, F_i ∈ R^L}
    ψ(s_i) ← [ (1/|R_k|) Σ_{j∈R_k} F_ij ]_{k=1}^K {region-level embedding}
end for
// Step 2: generate MSA subsets via Pattern-SF
{A'_1, ..., A'_M} ← PatternSelect(A, {ψ(s_i)}) {Mosaic-SF / Gradient-SF / Contrast-SF / Region-SF}
// Step 3: run structure predictor
D ← ∅
for m = 1 to M do
    for k = 1 to K do
        X_{m,k} ← f(A'_m, ξ_k)
        D ← D ∪ {X_{m,k}}
    end for
end for
// Step 4: coverage-aware refinement
Assign basin label ℓ(X) to each X ∈ D
for each basin b do
    D*_b ← TopK_pLDDT({X ∈ D : ℓ(X) = b}, K_b)
    D* ← D* ∪ D*_b
end for
return D*
```

SF-Cluster defines a subsampling map

$$\pi_{\mathrm{SF}} : \Phi(\mathcal{A}) \to \{\mathcal{A}'_1, \ldots, \mathcal{A}'_M\}, \tag{6}$$

which selects MSA subsets according to structure in frustration space rather than sequence similarity. Homologs with similar frustration profiles contribute a shared energetic context that can bias predictions toward particular conformational basins.

3.5. Pattern-SF: Practical Approximation via Pattern Geometry

Because $g_s(\cdot)$ is not directly observable, we approximate alignment to B_{s^*} using structured operations on a region-level compression of F_i. We partition residues into K contiguous regions $\mathcal{R} = \{R_k\}_{k=1}^K$ and define

$$\psi(\mathbf{s}_i) = [\mu_{i1}, \ldots, \mu_{iK}], \qquad \mu_{ik} = \frac{1}{|R_k|} \sum_{j \in R_k} F_{ij}, \tag{7}$$

where F_{ij} denotes the frustration index at residue j of homolog i. Subset selection is performed through a proxy objective in pattern space,

$$\mathcal{A}' = \arg\max_{\mathcal{A}' \subseteq \mathcal{A}} \mathcal{J}(\psi(\mathcal{A}')), \tag{8}$$

where $\mathcal{J}$ is assumed to be monotonically related to g_{s^*}.

We consider four strategies based on different geometric operations in ψ-space:

- **Mosaic-SF** selects homologs from multiple pattern modes, increasing variability in $\phi(\mathcal{A}')$ and weakening the dominance of a single frustration profile.
- **Gradient-SF** assigns each homolog a score
$$g_i = \sum_k w_k \mu_{ik}, \tag{9}$$
and selects subsets along the resulting directional axis.
- **Contrast-SF** selects homologs with large frustration differentials
$$c_i^{(a,b)} = \mu_{ia} - \mu_{ib}, \tag{10}$$
between specified regions, emphasizing localized energetic differences.
- **Region-SF** clusters homologs using k-means in the embedding space ψ, providing a frustration-space analogue of AF-Cluster.

Each strategy modifies the distribution of representations contained in $\mathcal{A}'$. Under the model of Section 3.2, these changes induce corresponding shifts in basin probabilities.

3.6. Coverage-Aware Refinement

Aggregating predictions across all subsets yields a candidate pool

$$\mathcal{D} = \bigcup_m \mathcal{D}(\mathcal{A}'_m). \tag{11}$$

Selecting structures solely by confidence corresponds to

$$\max_{\mathcal{D}'} \sum_{\mathbf{X} \in \mathcal{D}'} \mathrm{pLDDT}(\mathbf{X}), \tag{12}$$

which favors dominant conformational states and can eliminate lower-confidence rare states. We therefore impose a per-basin coverage constraint:

$$\mathcal{D}^* = \arg\max \sum_{\mathbf{X} \in \mathcal{D}'} \mathrm{pLDDT}(\mathbf{X}) \quad \text{s.t.} \quad |\mathcal{D}' \cap B_s| = K_s, \forall s, \tag{13}$$

where K_s denotes the quota assigned to basin B_s. This retains structurally rare predictions even when they have lower confidence scores, producing a coverage-balanced ensemble.

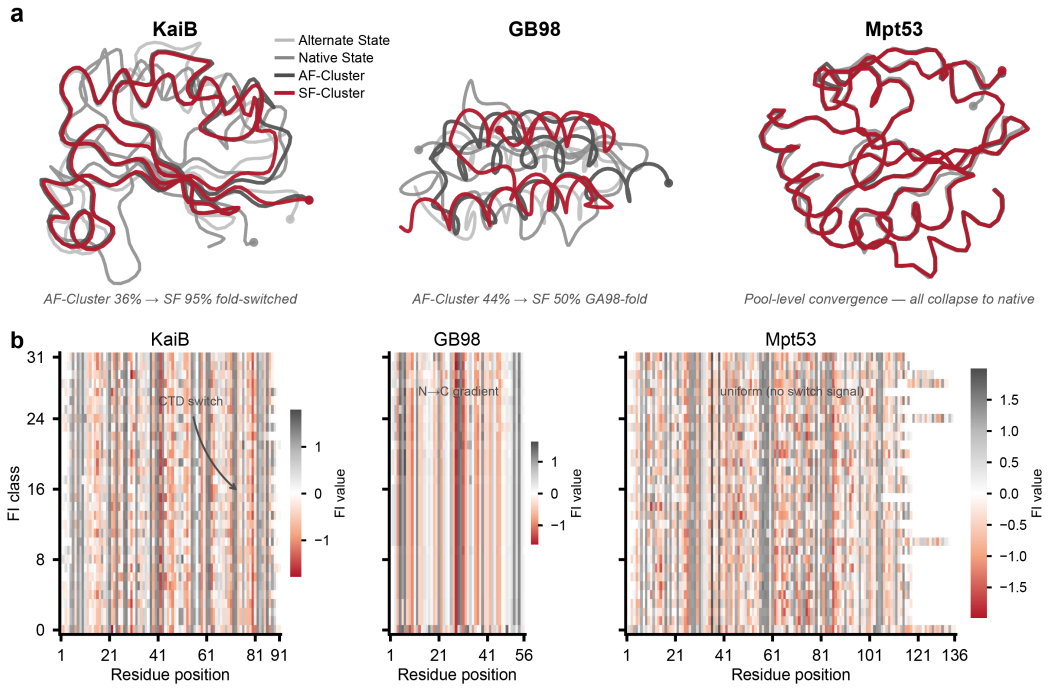

Figure 2. **Case-study visualization of structural outcomes and frustration maps.** Top: representative structural overlays for KaiB, GB98, and Mpt53. SF-Cluster strongly enriches the KaiB fold-switched state and partially enriches the GB98 target state, while all Mpt53 predictions collapse to the native basin. Bottom: residue-level frustration maps show that successful cases contain structured regional patterns, whereas Mpt53 lacks a productive switch signal.

3.7. Theoretical Analysis

The energy model in Section 3.2 yields two results characterizing the capabilities and limitations of frustration-guided subsampling. Full proofs are provided in Appendix A.

Theorem 1 (Subsampling as energy reweighting). Under the energy model, changing $\mathcal{A}'$ changes $\phi(\mathcal{A}')$, which induces a corresponding change in basin probabilities. The choice of representation therefore determines the extent to which subsampling can alter the conformational distribution.

Theorem 2 (Focusing and impossibility). *(a)* If there exists $\mathcal{A}'$ such that $g_{s^*}(\phi(\mathcal{A}')) \gg g_s(\phi(\mathcal{A}'))$ for all $s \neq s^*$, then $P(B_{s^*} \mid \mathcal{A}') \approx 1$, and targeted focusing is achievable.
(b) If $g_{s_0}(\phi(\mathcal{A}')) \geq g_s(\phi(\mathcal{A}'))$ for all $\mathcal{A}' \subseteq \mathcal{A}$ and all s, then no subsampling strategy can shift probability away from B_{s_0}. In this regime the MSA pool contains a single accessible basin, and alternative conformations cannot be recovered through subsampling alone.

Theorem 2(b) explains the behavior observed for Mpt53 and identifies a regime in which MSA subsampling cannot recover alternative conformational states.

Table 1. Targeted recovery on fold-switching benchmarks. Hit rate is the fraction of AF2 runs assigned to the target state. *GB98 T25I/L20A is a mutation-ablation control; see Section 4.4.

Protein	Strategy	AF-Cluster	SF-Cluster	RMSD (Å)
KaiB	Mosaic-SF	36.2%	**95.0%**	2.88
GA98	Mosaic-SF	46.2%	**92.5%**	0.70
GB98	Gradient-SF	43.8%	**50.0%**	0.74
GB98*	-	45.0%	0.0%	-

3.8. Overall Algorithm

Algorithm 1 summarizes the SF-Cluster pipeline. Frustration profiles are computed for all homologs and compressed into region-level embeddings. Pattern-SF generates MSA subsets with distinct frustration-space structure, and coverage-aware refinement assembles the final ensemble while preserving representation across conformational basins.

4. Experiments

4.1. Experimental Setup

We evaluated SF-Cluster on fold-switching benchmarks (KaiB, GA98, and GB98), a discovery benchmark (Mpt53), and held-out proteins (RfaH and MAD2).

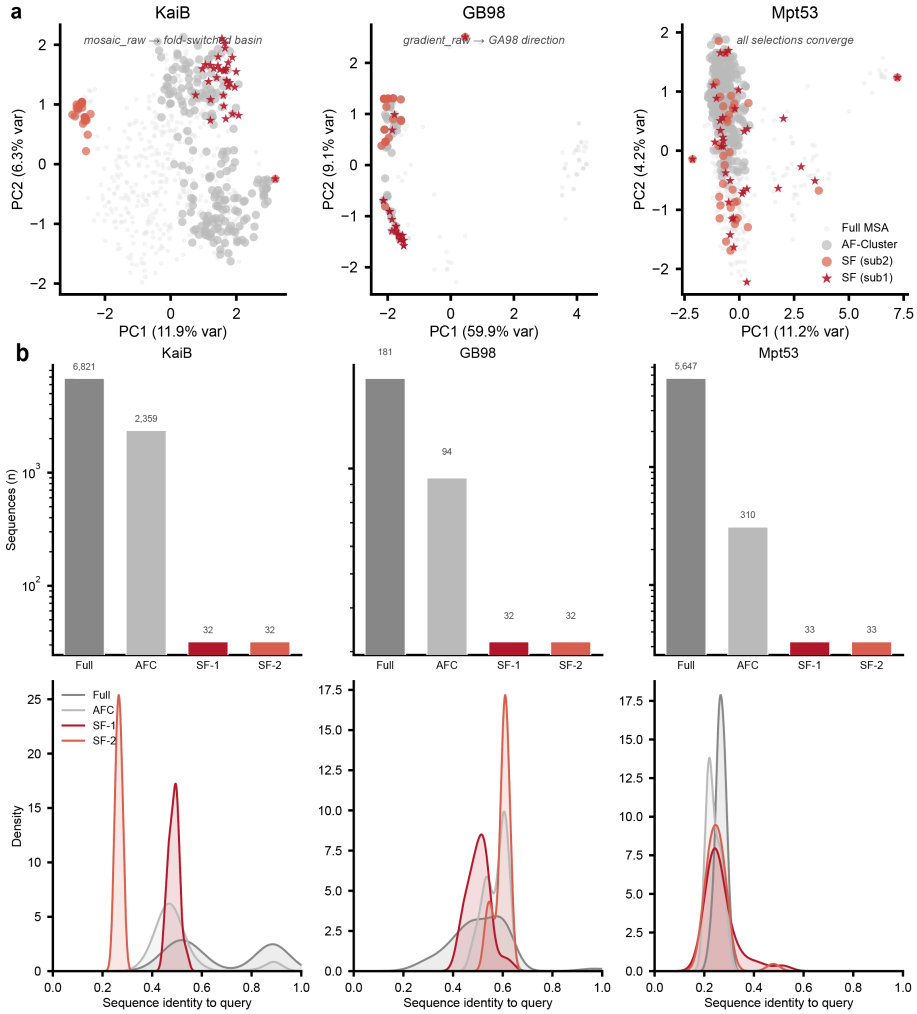

Figure 3. **MSA subset geometry underlying SF-Cluster behavior.** Top: projection of MSA sequences into frustration-pattern space. Successful SF subsets occupy productive directions for KaiB and GB98, whereas Mpt53 selections converge to the same basin despite occupying different regions of sequence space. Middle and bottom: MSA-depth and sequence-identity distributions show that SF-Cluster does not merely select larger or more similar subsets; it selects compact subspaces with distinct pattern geometry.

Table 2. Target-state hit rate by Pattern-SF strategy. Column-wise best is bolded. No single strategy dominates across proteins; all strategies collapse to the reference basin on Mpt53.

Strategy	KaiB	GA98	GB98
Mosaic-SF	**95.0%**	**92.5%**	18.8%
Gradient-SF	65.0%	50.0%	**50.0%**
Contrast-SF	67.5%	45.0%	7.5%
Region-SF	53.8%	50.0%	45.0%
AF-Cluster	36.2%	46.2%	43.8%

Table 3. Basin analysis on Mpt53. A collapse score near 1.0 indicates convergence to the reference state; the lower score for AF-Cluster reflects structural diversity without non-reference basin recovery. All SF-Cluster variants not shown produce identical outcomes.

Method	Collapse	TM	Alt. basin
AF-Cluster	0.21	0.622	No
Random	1.00	0.971	No
Mosaic-SF	1.00	0.970	No
Gradient-SF	1.00	0.970	No

Datasets. KaiB, GA98, GB98, and Mpt53 form the development set, while RfaH and MAD2 are reserved for external validation. GB98 T25I/L20A serves as a negative control in which two substitutions ablate the fold-switch signal without substantially altering MSA composition. Additional dataset details are provided in Appendix B.1.

Table 4. Rare-state hits retained after refinement. *GB98 T25I/L20A; see Table 1.

Protein	Variant	Global pLDDT	Coverage-aware
KaiB	KaiB	0 hits	**81** hits
GA/GB	GA98	77 hits	37 hits
	GB98	70 hits	35 hits
	GB98*	36 hits	36 hits

Table 5. Signal attribution controls grouped by degree of positional coupling preserved. Mosaic-SF and Gradient-SF rows here use residue-normalized FI values and are included as ablation controls; their raw-FI counterparts are the deployable variants reported in Table 2.

Control	Coupling	GA98	GB98
Frust-Outlier	full	**98.8%**	36.3%
Frust-Diverse	full	96.3%	**50.0%**
Frust-Balanced	full	61.3%	**50.0%**
Frust-Local	partial	75.0%	47.5%
Switch-column	partial	75.0%	50.0%
Shuffled-FI	none	72.5%	66.3%
Random-column	none	97.5%	53.0%
Mosaic-SF (res)	full, normalized	83.8%	5.0%
Gradient-SF (res)	full, normalized	46.3%	8.8%

Baselines and variants. We compared SF-Cluster against full-MSA prediction and AF-Cluster (Wayment-Steele et al., 2024). Pattern-SF variants (Mosaic-SF, Gradient-SF, Contrast-SF, and Region-SF) operate on raw frustration index values, while Section 4.6 examines residue-level normalization as an ablation. All methods use the same AlphaFold2 backend and screen-and-refine protocol (Appendix B.6).

Evaluation metrics. For fold-switching proteins, we measured target-state hit rate using switch-region RMSD together with a uniform pLDDT threshold. For Mpt53, we report collapse score and best-hit RMSD. Complete metric definitions are provided in Appendix B.8.

4.2. SF-Cluster Improves Targeted Conformational Focusing

We first evaluated whether frustration-guided subsampling improves recovery of known alternative conformations relative to sequence-space clustering (Table 1). For KaiB, Mosaic-SF recovered the fold-switched state in 95.0% of predictions compared with 36.2% for AF-Cluster. For GA98, Mosaic-SF reached 92.5%, approximately doubling the AF-Cluster hit rate. GB98 proved more challenging: only Gradient-SF improved upon AF-Cluster, reaching a hit rate of 50.0%.

These results indicate that modifying the frustration composition of an MSA can shift the conformational distribution sampled during structure prediction and improve recovery of specific target states.

4.3. Pattern Geometry Determines Sampling Effectiveness

We next asked whether a single frustration-based sampling strategy performs consistently across proteins (Table 2). Mosaic-SF achieved the highest hit rates for KaiB and GA98, indicating that the relevant conformational signal is distributed across multiple frustration regions. In contrast, Gradient-SF performed best on GB98, consistent with a signal concentrated along a directional N-to-C frustration gradient. Region-SF consistently underperformed Mosaic-SF and Gradient-SF across all systems. These results suggest that effective subsampling depends on the structure of frustration space within each protein.

KaiB exhibits a broad multi-region frustration pattern captured by Mosaic-SF, whereas GB98 contains a directional gradient tracked by Gradient-SF (Figure 2). Mpt53 shows little regional organization in its frustration profile, consistent with a single-basin MSA pool. Successful SF-Cluster subsets occupy distinct regions of frustration space (Figure 3) and are not simply larger or more query-similar subsets.

4.4. MSA Subsampling Cannot Create Missing Conformational Basins

We next examined the failure mode predicted by Theorem 2(b). Mpt53 provides a setting in which the MSA appears to encode only a single structural basin. Across AF-Cluster, random sampling, and all SF-Cluster variants, we did not recover any non-reference basin (Table 3). All SF-Cluster variants collapsed to the reference structure. AF-Cluster produced greater structural diversity, reflected by its lower collapse score, but none of these structures corresponded to a validated alternative basin.

The GB98 T25I/L20A control provides a complementary perspective. AF-Cluster continued to recover the GA98-like fold in 45.0% of predictions, whereas all SF-Cluster variants failed entirely. Removing the frustration-pattern signal therefore eliminates targeted recovery even though sequence-space exploration still accesses the alternative fold. Together, these experiments indicate that subsampling can enrich conformational states already represented within an MSA but cannot recover states absent from the underlying sequence pool.

Table 6. External validation on RfaH and MAD2. Target hit rate is the fraction of predictions assigned to the minority conformation. RfaH AF-Cluster was run with a memory-capped MSA; its median pLDDT of 59.4 falls below the hit confidence threshold.

Protein	Method	Hit Rate	Median pLDDT	Best RMSD (Å)	Note
RfaH	Full-MSA	0.0%	74.5	11.36	Autoinhibited state only
	AF-Cluster†	4.3%	59.4	2.17	Below confidence threshold
	SF-Cluster	**29.2%**	**78.3**	2.28	Best arm (Frust-Outlier)
MAD2	Full-MSA	**100.0%**	**91.1**	0.97	Ceiling performance
	AF-Cluster	3.6%	58.6	1.23	Underperforms full-MSA
	SF-Cluster	75.0%	85.1	**0.87**	Best arm (Frust-Diverse)

† RfaH AF-Cluster was run with capped MSA depth to avoid GPU memory failure; the biological evaluation protocol was unchanged.

4.5. Coverage-Aware Refinement Prevents Confidence-Driven Loss of Rare States

We compared global pLDDT ranking against coverage-aware refinement (Table 4). For KaiB, global pLDDT ranking retained no fold-switched predictions, whereas coverage-aware refinement retained 81. The fold-switched state consistently received lower confidence scores than the dominant conformation, causing confidence-based selection to remove it from the final ensemble.

GA98 and GB98 exhibited different behavior. Because the target states achieved pLDDT values comparable to the dominant conformations, both refinement procedures produced similar results. Coverage-aware refinement therefore becomes most important when the target conformation systematically receives lower confidence scores than the dominant state.

4.6. Signal Attribution

To identify which aspects of the frustration representation contribute to performance, we evaluated controls that preserve or destroy different forms of structure in the frustration signal. Frust-Outlier, Frust-Diverse, and Frust-Balanced preserve full positional coupling while sampling the frustration distribution differently. Frust-Local and Switch-column retain only partial positional structure. Shuffled-FI preserves per-position marginals while destroying inter-position coupling, and Random-column removes frustration information entirely.

Methods preserving positional coupling consistently achieved the highest hit rates (Table 5). On GA98, Shuffled-FI still achieved a 72.5% hit rate and Random-column reached 97.5%, indicating that sequence composition alone carries substantial information about the target state. GB98 was considerably more sensitive to disruption of the frustration signal. Random-column dropped to 53.0% and Shuffled-FI to 66.3%, while methods preserving positional coupling remained more consistent. Residue-normalized variants also underperformed their raw counterparts, suggesting that absolute frustration magnitudes contain information beyond relative patterns alone.

4.7. External Validation

On RfaH, SF-Cluster recovered the fold-switched CTD conformation in 29.2% of predictions with a best-hit RMSD of 2.28 Å (Table 6), whereas full-MSA prediction failed to recover the target state. AF-Cluster achieved 4.3% recovery, although its median pLDDT of 59.4 fell below the confidence threshold applied throughout this study. Relative to this baseline, SF-Cluster improved hit rate by 6.8× while using approximately half the prediction budget.

On MAD2, full-MSA prediction already recovered the target state at ceiling performance. SF-Cluster reached 75.0% recovery but did not improve upon the full-MSA baseline, and neither method recovered the open O-MAD2 state. These results indicate that SF-Cluster generalizes when the MSA contains a usable conformational signal but provides limited benefit when a dominant state already saturates performance.

5. Conclusion

Here, we introduce SF-Cluster, a frustration-aware MSA subsampling framework that guides AlphaFold predictions through frustration-pattern representations derived from homologous sequences. Across fold-switching benchmarks, SF-Cluster improves targeted recovery of alternative conformations and shows that conformational information encoded within an MSA extends beyond sequence similarity alone. Our results indicate that different proteins favor different frustration-space sampling strategies, that coverage-aware refinement is often necessary to preserve low-frequency states, and that MSA subsampling operates through representation-induced reweighting of conformational probabilities. The Mpt53 and GB98 T25I/L20A experiments further establish that the conformational diversity accessible through subsampling depends on the information contained within the underlying MSA pool. Together, these findings identify frustration patterns as a useful representation for controllable conformational sampling and highlight MSA representation as a powerful mechanism for influencing the conformational states recovered by modern structure prediction systems.

Impact Statement

This work advances the controllable prediction of alternative protein conformations, which has potential benefits for understanding allosteric regulation, fold-switching, and other functionally important structural transitions relevant to drug discovery and basic biology. By framing MSA subsampling as a focusing rather than a discovery mechanism, SF-Cluster also clarifies the limits of what current sequence-based methods can achieve, which may help practitioners avoid overinterpreting predicted conformational ensembles. We do not anticipate direct negative societal consequences beyond those generally associated with protein structure prediction, though we note that any computational tool capable of characterizing conformational states could in principle be misapplied to engineer proteins with harmful properties; the method here neither lowers existing barriers to such misuse nor introduces new ones, as it reweights conformations already encoded in natural homolog pools rather than designing novel structures.

Acknowledgment

This research was supported by a grant from the Highthroughput Institute for Discovery (HIT-ID) at the University of Pennsylvania to the lab of Pranam Chatterjee. The work described in this paper was also partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N.

References

Bryant, P. and Noé, F. Structure prediction of alternative protein conformations. *Nature Communications*, 15(1): 7328, 2024.

Cao, H., Zhou, X., Gao, Z., Wang, C., Gao, X., Zhang, Z., Gu, C., Liu, G., and Heng, P.-A. Plame: Lightweight msa design advances protein folding from evolutionary embeddings. In *NeurIPS 2025 AI for Science Workshop*, 2025.

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

Chakravarty, D., Schafer, J. W., Chen, E. A., Thole, J. F., Ronish, L. A., Lee, M., and Porter, L. L. Alphafold predictions of fold-switched conformations are driven by structure memorization. *Nature communications*, 15(1): 7296, 2024.

Chen, B., Bei, Z., Cheng, X., Li, P., Tang, J., and Song, L. Msagpt: Neural prompting protein structure prediction via msa generative pre-training. *Advances in Neural Information Processing Systems*, 37:37504–37534, 2024.

Del Alamo, D., Sala, D., Mchaourab, H. S., and Meiler, J. Sampling alternative conformational states of transporters and receptors with alphafold2. *elife*, 11:e75751, 2022.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Lee, M., Schafer, J. W., Prabakaran, J., Chakravarty, D., Clore, M. F., and Porter, L. L. Large-scale predictions of alternative protein conformations by alphafold2-based sequence association. *Nature Communications*, 16(1): 5622, 2025.

Leusch, J.-P., Poley-Gil, M., Fernandez-Martin, M., Bordin, N., Rost, B., Parra, R. G., and Heinzinger, M. Frustraiseq: Scaling local energetic frustration to the protein sequence space. *bioRxiv*, pp. 2026–02, 2026.

Li, S., Zhang, C., Kong, L., Xue, Y., Liu, S., and Gao, Y. Q. Disentangling coevolutionary constraints for modeling protein conformational heterogeneity. *Communications Chemistry*, 9:146, 2026.

Sgarbossa, D., Lupo, U., and Bitbol, A.-F. Generative power of a protein language model trained on multiple sequence alignments. *Elife*, 12:e79854, 2023.

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

Venkatraman, A., Cao, H., Wei, T., Cheng, C., and Liu, G. Msaflow: a unified approach for msa representation, augmentation, and family-based protein design. In *NeurIPS 2025 AI for Science Workshop*.

Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., and Kern, D. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, 625(7996): 832–839, 2024.

Xing, E., Zhang, J., Wang, S., and Cheng, X. Leveraging sequence purification for accurate prediction of multiple conformational states with alphafold2. *Research Square*, pp. rs–3, 2025.

Zhang, J., Liu, S., Chen, M., Chu, H., Wang, M., Wang, Z., Yu, J., Ni, N., Yu, F., Chen, D., et al. Unsupervisedly prompting alphafold2 for accurate few-shot protein structure prediction. *Journal of Chemical Theory and Computation*, 19(22):8460–8471, 2023.

Zhang, L., Chen, J., Shen, T., Li, Y., and Sun, S. Msa generation with seqs2seqs pretraining: advancing protein structure predictions. *Advances in Neural Information Processing Systems*, 37:57324–57348, 2024.

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

A. Theoretical Analysis

A.1. Setup

Let $\mathcal{A}$ be a full MSA and $\mathcal{A}' \subseteq \mathcal{A}$ a selected sub-MSA. A structure predictor induces a conditional distribution $P(\mathbf{X} \mid \mathcal{A}')$ over structure space, partitioned into conformational basins $\mathcal{B} = \{B_1, \ldots, B_S\}$ with basin probabilities

$$P(B_s \mid \mathcal{A}') = \int_{B_s} P(\mathbf{X} \mid \mathcal{A}')\, d\mathbf{X}.$$

We assume the predictor depends on the MSA through a representation $\phi(\mathcal{A}')$, and that basin probabilities take the softmax form

$$P(B_s \mid \mathcal{A}') = \frac{\exp(g_s(\phi(\mathcal{A}')))}{\sum_{r=1}^{S} \exp(g_r(\phi(\mathcal{A}')))},$$

where $g_s(\cdot)$ is an unknown basin-specific compatibility score.

A.2. Theorem 1: MSA Subsampling Induces Conformational Reweighting

Theorem 1. *Suppose there exist two sub-MSAs $\mathcal{A}'_1, \mathcal{A}'_2 \subseteq \mathcal{A}$ such that $\phi(\mathcal{A}'_1) \neq \phi(\mathcal{A}'_2)$, and for some basin B_s and at least one $r \neq s$,*

$$\big[g_s(\phi(\mathcal{A}'_1)) - g_r(\phi(\mathcal{A}'_1))\big] \;\neq\; \big[g_s(\phi(\mathcal{A}'_2)) - g_r(\phi(\mathcal{A}'_2))\big].$$

Then $P(B_s \mid \mathcal{A}'_1) \neq P(B_s \mid \mathcal{A}'_2)$.

Proof. The log-odds between basins s and r under any sub-MSA $\mathcal{A}'$ is

$$\log \frac{P(B_s \mid \mathcal{A}')}{P(B_r \mid \mathcal{A}')} = g_s(\phi(\mathcal{A}')) - g_r(\phi(\mathcal{A}')).$$

By assumption, this quantity differs between $\mathcal{A}'_1$ and $\mathcal{A}'_2$ for at least one $r \neq s$. Hence the relative odds of B_s versus B_r change between the two sub-MSAs, which implies $P(B_s \mid \mathcal{A}'_1) \neq P(B_s \mid \mathcal{A}'_2)$. □

Theorem 1 establishes that MSA subsampling is a conformational reweighting operation: its effect is entirely mediated by the shift in representation $\phi(\mathcal{A}')$, making the choice of ϕ the binding constraint on what subsampling can achieve.

A.3. Theorem 2: Conditions for Focusing and Impossibility

Theorem 2. ***(a) Targeted focusing.*** *Suppose there exists a target basin B_{s^*} and a sub-MSA $\mathcal{A}'$ such that*

$$g_{s^*}(\phi(\mathcal{A}')) - g_s(\phi(\mathcal{A}')) \;\geq\; \Delta \quad \forall\, s \neq s^*.$$

Then $P(B_{s^*} \mid \mathcal{A}') \;\geq\; \frac{1}{1+(S-1)e^{-\Delta}}$, and in particular $P(B_{s^*} \mid \mathcal{A}') \to 1$ as $\Delta \to \infty$.

(b) Single-basin impossibility. *Suppose there exists a dominant basin B_{s_0} such that for every $\mathcal{A}' \subseteq \mathcal{A}$,*

$$g_{s_0}(\phi(\mathcal{A}')) - g_s(\phi(\mathcal{A}')) \;\geq\; \Delta \quad \forall\, s \neq s_0.$$

Then for every sub-MSA $\mathcal{A}'$, $P(B_{s_0} \mid \mathcal{A}') \;\geq\; \frac{1}{1+(S-1)e^{-\Delta}}$, and no row-subsetting strategy can recover an alternative basin with high probability when Δ is large.

Proof of Part (a). Let $z_s = g_s(\phi(\mathcal{A}'))$. By assumption, $z_s \leq z_{s^*} - \Delta$ for all $s \neq s^*$. Then $\sum_{s \neq s^*} e^{z_s} \leq (S-1)e^{z_{s^*}} e^{-\Delta}$. Substituting into the softmax denominator,

$$P(B_{s^*} \mid \mathcal{A}') = \frac{e^{z_{s^*}}}{e^{z_{s^*}} + \sum_{s \neq s^*} e^{z_s}} \;\geq\; \frac{e^{z_{s^*}}}{e^{z_{s^*}} + (S-1)e^{z_{s^*}} e^{-\Delta}} = \frac{1}{1 + (S-1)e^{-\Delta}}.$$

As $\Delta \to \infty$, $e^{-\Delta} \to 0$, giving $P(B_{s^*} \mid \mathcal{A}') \to 1$. □

Proof of Part (b). The proof is identical to Part (a) with s^* replaced by s_0, but the margin condition holds uniformly over all $\mathcal{A}' \subseteq \mathcal{A}$. For any selected sub-MSA, the same bound gives $P(B_{s_0} \mid \mathcal{A}') \geq \frac{1}{1+(S-1)e^{-\Delta}}$. Since this applies to every feasible subset, no subsampling strategy can meaningfully shift probability mass away from B_{s_0}. □

Implication. Part (b) formalises the empirical observation that SF-Cluster fails to recover alternative conformations for Mpt53: when the MSA pool is structurally single-basin, every feasible $\phi(\mathcal{A}')$ remains dominated by g_{s_0}, and subsampling cannot create conformations absent from the input pool.

B. Implementation Details

B.1. Protein Systems and Reference Structures

We study six proteins spanning three experimental categories. KaiB, GA98, GB98, and Mpt53 constitute the development set; RfaH and MAD2 are held out for external validation. GB98 T25I/L20A is a designed negative control derived from the GB98 sequence and is not used for method development.

Table 7. Reference structures for all protein systems. State A denotes the dominant or native conformation; State B denotes the fold-switched or target alternative. Construct lengths correspond to the residue ranges used for all predictions and evaluations.

Protein	Length	State A	State B	UniProt
KaiB	91 aa	2QKE (chain B)	5JYT (chain A)	Q79V61
GA98	56 aa	2LHD (chain A)	2LHC (chain A)	-
GB98	56 aa	2LHD (chain A)	2LHC (chain A)	-
GB98 T25I/L20A	56 aa	2LHE (chain A)	-	-
Mpt53	136 aa	1LU4 (chain A)	discovery task	P9WG65
RfaH	162 aa	5OND (chain A)	6C6S (chain D)	P0AFW0
MAD2	205 aa	1DUJ (chain A)	1GO4 (chain A)	Q13257

KaiB. The construct spans residues 5-95 of PDB 2QKE chain B, trimmed from the 108-residue crystal chain to remove disordered terminal residues. State B (5JYT) carries stabilizing mutations Y8A, N29A, G89A, D91R, and Y94A relative to wild-type, and was determined by NMR as an ensemble of 20 models; the first model is used as the reference.

GA/GB proteins. The GA and GB protein families are combinatorially engineered proteins of 56 residues that adopt distinct α-helical (GA fold) and β-sheet (GB fold) structures. The primary fold-switch pair evaluated is GB98 (PDB 2LHD, GB fold) as state A versus GA98 (PDB 2LHC, GA fold) as state B. Structures were retrieved from the RCSB and trimmed to residues 1-56; no further trimming was required.

Mpt53. The construct spans residues 38-173 in UniProt coordinates (P9WG65), corresponding to crystal residues 1001-1134 of PDB 1LU4 chain A with offset 963 applied. Because no experimentally confirmed alternative fold is known for Mpt53, this protein is treated as a discovery benchmark: the objective is to identify non-native structural basins rather than recover a known target state.

RfaH. Full-length RfaH (residues 1-162) is used. State A is the NusG-like autoinhibited conformation (5OND) and state B is the fold-switched CTD conformation observed in the RNAP elongation complex (6C6S). The fold-switch region corresponds to the C-terminal domain, approximately residues 101-162.

MAD2. The full 205-residue construct is used. State A is the open O-MAD2 conformation (1DUJ) and state B is the closed C-MAD2 conformation in complex with MAD1 (1GO4). The conformational switch involves topological rearrangement of the safety-belt region spanning approximately residues 170-205.

GB98 T25I/L20A. This double mutant of GB98 (PDB 2LHE) is used exclusively as a negative control. The two substitutions occur at positions implicated in the fold-switch interface, together ablating the fold-switch signal detectable in the MSA. It is not used for method development or tuning.

B.2. MSA Construction

Multiple sequence alignments were generated using ColabFold v1.6.1 with the `-msa-only` flag. The MMseqs2 search was performed against the combined UniRef30 and environmental sequence databases via the public ColabFold API server. All queries were submitted as single FASTA sequences without template search. MSAs were retrieved in A3M format.

Raw A3M files were filtered following the protocol of Wayment-Steele et al. (2024): a non-query sequence is retained if and only if the fraction of gap characters over aligned columns is at most 25%. Lowercase A3M insertion characters are excluded from the gap count. The query sequence is always retained unconditionally.

Table 8. MSA statistics after gap filtering. Neff is computed at a sequence identity threshold of 0.8 using the standard reweighting formula; it is not reported for external validation cases as these MSAs were not used for method development. Neff is not reported for RfaH and MAD2 as these are validation-only cases.

Protein	Variant	Raw depth	Filtered	Length	Neff
KaiB	KaiB	7,895	6,821	91	4,570
GA/GB	GA98	438	179	56	22.1
	GB98	396	181	56	26.1
	GB98 T25I/L20A	369	169	56	26.6
Mpt53	Mpt53	8,418	5,647	136	4,801
RfaH	RfaH	5,606	3,359	162	-
MAD2	MAD2	3,179	2,091	205	-

B.3. Frustration Index Computation

Per-residue frustration indices are computed using FrustrAI-Seq (Leusch et al., 2026), a sequence-based frustration predictor built on a ProtT5-XL encoder backbone. For each homolog in the filtered MSA, the ungapped sequence is extracted and passed to the model, which returns a per-residue frustration index (FI), a class label, and a surprisal score. Sequences containing non-canonical amino acids are excluded and filled with NaN in all downstream matrices.

The resulting per-residue vectors are projected back to alignment coordinates using each sequence's uppercase-to-column mapping from the A3M format, yielding FI, entropy, and surprisal matrices of shape $N \times L$, where N is the number of homologs and L is the alignment length. Gap positions receive NaN. Inference runs with batch size 1 on a single NVIDIA A100-SXM4-80GB GPU under `torch.no_grad()`.

To identify positions of high evolutionary and energetic variation, each alignment column c is scored by

$$s_c = \mathrm{Var}(\mathrm{FI}_c) \cdot (1 - \overline{\mathrm{entropy}}_c) \cdot (1 + \overline{|\mathbf{z}|}_c), \tag{14}$$

where $\mathrm{Var}(\mathrm{FI}_c)$ is the across-sequence variance of FI at column c, $\overline{\mathrm{entropy}}_c$ is the mean per-sequence entropy, and $\overline{|\mathbf{z}|}_c$ is the mean absolute surprisal. The top-30 columns by this score are retained as candidate switch columns for downstream use in the outlier-guided arm.

B.4. Pattern-Level Feature Representation

Region-level embeddings compress each homolog's FI profile into a fixed-dimensional feature vector. Two complementary segmentation schemes are applied to the alignment columns:

Structural segmentation divides columns into an N-terminal half (columns 1 to $\lfloor L/2 \rfloor$) and a C-terminal half (columns $\lfloor L/2 \rfloor + 1$ to L).

Variance-based segmentation designates the top 20% of columns by across-sequence FI variance as high-variance positions and the remaining 80% as low-variance positions.

Per-sequence features are then computed from these partitions:

Two FI weighting variants are used: raw FI values and residue-normalized FI values in which the per-column mean is subtracted before aggregation. The raw variant is used in all main results; the normalized variant is evaluated as an ablation in Section 4.6.

B.5. Subsampling Strategies

All subsampling strategies produce 12 subsets of 32 sequences each per arm. The query sequence is always prepended to every subset. Random seeds are fixed at 0 throughout.

Table 9. Per-sequence features derived from the frustration index matrix. All means and variances exclude gap and invalid positions. Feature vectors are standardized (zero mean, unit variance) before clustering.

Feature	Definition
$\bar{\mathrm{F}}\mathrm{I}_{\mathrm{N}}$	Mean FI over N-terminal half
$\bar{\mathrm{F}}\mathrm{I}_{\mathrm{C}}$	Mean FI over C-terminal half
$\bar{\mathrm{F}}\mathrm{I}_{\mathrm{hv}}$	Mean FI over high-variance columns
$\bar{\mathrm{F}}\mathrm{I}_{\mathrm{lv}}$	Mean FI over low-variance columns
Δ_{NC}	$\bar{\mathrm{F}}\mathrm{I}_{\mathrm{N}} - \bar{\mathrm{F}}\mathrm{I}_{\mathrm{C}}$
Δ_{hvlv}	$\bar{\mathrm{F}}\mathrm{I}_{\mathrm{hv}} - \bar{\mathrm{F}}\mathrm{I}_{\mathrm{lv}}$
σ^2_{N}	Within-sequence FI variance over N-terminal half
σ^2_{C}	Within-sequence FI variance over C-terminal half

AF-Cluster. The filtered MSA is partitioned using DBSCAN on one-hot encoded sequences over the 21-character alphabet (20 amino acids plus gap), yielding feature vectors of dimension $21L$. The ϵ parameter is selected by scanning $[3.0, 20.0]$ in steps of 0.5, applying DBSCAN at each step to a random 25% subsample; the ϵ maximizing the number of clusters is used. Final clustering runs on the full MSA at the selected ϵ, with `min_samples` $= 3$. Noise points are discarded.

Arm B (F-Cluster). Each sequence is embedded by its L2-normalized FI profile, with gap positions imputed to 0. k-means with $k = 12$ is applied; within each cluster, the 32 sequences nearest the centroid in embedding space are selected.

Arm C (Hybrid-Cluster). The pairwise distance used for clustering is

$$d_{ij} = 0.5\, d_{\text{Hamming}}(i,j) + 0.3\, d^{\text{FI}}_{\cos}(i,j) + 0.2\, d^{\text{entropy}}_{\cos}(i,j), \tag{15}$$

where d_{Hamming} is the normalized pairwise Hamming distance on aligned columns and $d_{\cos}$ is the cosine dissimilarity in the respective feature space. For MSAs with $N > 3{,}000$ sequences, the pairwise Hamming matrix is approximated by concatenating FI and entropy feature vectors weighted by their respective coefficients, and k-means is applied directly. For smaller MSAs, the distance matrix is embedded into 10-dimensional Euclidean space via classical multidimensional scaling before k-means clustering.

Arm D (Outlier-guided). Each sequence is scored over the top-30 switch columns:

$$s_i = \sum_{c \in \mathcal{S}} |z_{i,c}| \cdot (1 - \text{entropy}_{i,c}), \tag{16}$$

where $z_{i,c}$ is the surprisal at position c and gap entries contribute 0. Sequences are sorted by s_i descending and partitioned into 12 bands via round-robin assignment.

Arm E (Diversity-guided). Twelve subset seeds are chosen by farthest-point sampling in L2-normalized FI space. Each subset consists of the seed and its 31 nearest neighbors in that space. This arm maximizes representational diversity across subsets, complementing the compactness of Arm B.

Phase XII Pattern-SF arms. Four strategies operate on the region-level feature vectors defined in Section B.4:

- **Region-cluster** applies k-means ($k = 12$) to the six-dimensional vector ($\bar{\mathrm{F}}\mathrm{I}_{\mathrm{N}}$, $\bar{\mathrm{F}}\mathrm{I}_{\mathrm{C}}$, $\bar{\mathrm{F}}\mathrm{I}_{\mathrm{hv}}$, $\bar{\mathrm{F}}\mathrm{I}_{\mathrm{lv}}$, σ^2_{N}, σ^2_{C}) after standardization. Within each cluster, sequences are subsampled to 32.
- **Contrast** sorts sequences by $\Delta_{\mathrm{NC}} = \bar{\mathrm{F}}\mathrm{I}_{\mathrm{N}} - \bar{\mathrm{F}}\mathrm{I}_{\mathrm{C}}$ and partitions the sorted list into 12 equal bands, each forming one subset.
- **Mosaic** constructs each subset by strided sampling across the contrast-sorted list, mixing sequences with high and low Δ_{NC} within each subset to increase N/C frustration variance.
- **Gradient** sorts sequences by $\bar{\mathrm{F}}\mathrm{I}_{\mathrm{N}}$ and partitions into 12 bands, producing a directional sweep from N-frustrated to N-non-frustrated homologs.

B.6. Structure Prediction Protocol

All structure predictions use AlphaFold 2 via ColabFold batch interface (v1.6.1, `alphafold2_ptm` model). No structural relaxation is performed; pLDDT scores are read from the B-factor column of the output PDB. Predictions use custom A3M inputs only, without paired or environmental MSA supplement.

Predictions follow a two-stage screen-and-refine protocol. In the **screen stage**, all n subsets for an arm are run with 1 model and 2 seeds per subset, producing 2 structures per subset. In the **refine stage**, the top-K subsets selected by the refinement policy are re-run with 5 models and 4 seeds, producing 20 structures per subset. Both stages use 3 recycling iterations. The number of promoted subsets is

$$K = \max(4,\ \min(16,\ \lceil 0.2 \times n_{\text{subsets}} \rceil))\,. \tag{17}$$

The full-MSA baseline uses 5 models and 2 seeds with 3 recycling iterations, producing 10 structures per case.

B.7. Refinement Strategy

Global pLDDT selection. Each subset is represented by its screen-stage structure with the highest mean pLDDT. The top-K subsets by this score are promoted to the refine stage.

Coverage-aware refinement for fold-switching cases. For proteins with known alternative states (KaiB, GA98, GB98), each subset is assigned to the state with the lower switch-region RMSD among its screen structures. A subset is labelled ambiguous if $|\text{RMSD}_A - \text{RMSD}_B| < 1.0$ Å. The promotion budget is split equally across non-empty state partitions; unfilled quota is redistributed to the highest-pLDDT subsets in the remaining pool.

Per-basin refinement for discovery cases. For Mpt53, screen-stage representative structures are clustered by TM-score using single-linkage at a threshold of 0.80. Within each structural basin, the top $\lceil K/n_{\text{basins}} \rceil$ subsets by pLDDT are promoted. This procedure identifies structurally distinct prediction clusters without assuming prior knowledge of alternative states.

B.8. Evaluation Metrics

Structural evaluation regions. For each fold-switching case, two reference regions are pre-computed by pairwise TM-align superposition of the two reference structures, followed by DSSP secondary structure assignment. The **common core** consists of Cα atoms present in both states with per-residue displacement < 1.5 Å after superposition, excluding the 10 most mobile terminal residues on each side. The **switch region** consists of residues with displacement ≥ 3.0 Å or a change in secondary structure type (H/E/C) between the two states.

Table 10. Evaluation region sizes for fold-switching cases.

Protein	Common core	Switch region
KaiB	27 residues	30 residues
GA/GB	11 residues	20 residues

Cα RMSD. RMSD is computed after least-squares superposition on the common core residue set using BioPython 1.87. Predicted residues are renumbered to match the reference PDB's residue numbering before superposition. A minimum of 3 residues must be present in both the prediction and reference; otherwise RMSD is reported as NaN. The RMSD reported in all tables is computed over the common core.

TM-score. TM-score is computed using TMalign v20220412 (Zhang & Skolnick, 2005), run without threading options. Both TM1 (normalized by predicted chain length) and TM2 (normalized by reference chain length) are recorded.

Hit criterion. A prediction is classified as a hit for reference state S if and only if all three of the following conditions hold: (1) common-core RMSD ≤ 3.0 Å after superposition on the common core; (2) mean pLDDT over the full structure ≥ 70; (3) mean pLDDT over the switch region ≥ 70, where a switch region is defined. This criterion is applied identically to all methods.

State assignment. For fold-switching cases, each structure is assigned to the state with the lower switch-region RMSD. Structures where $|\mathrm{RMSD}_A - \mathrm{RMSD}_B| < 1.0$ Å are labelled ambiguous and excluded from hit counting. The hit rate reported in all tables is computed over non-ambiguous predictions.

Collapse score. For the Mpt53 discovery case, the collapse score of a prediction set $\mathcal{P}$ is defined as

$$\mathrm{collapse}(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{1}[\mathrm{TM}(p,\, 1\mathrm{LU4}) > 0.80], \tag{18}$$

where $\mathrm{TM}(p, 1\mathrm{LU4})$ is the TM-score of prediction p against the native reference. A collapse score of 1.0 indicates that all predictions lie within the native basin.

B.9. Hyperparameters

All hyperparameters are fixed before any AlphaFold 2 inference is run. The table below lists all parameters that materially affect results.

Table 11. Hyperparameter settings. All values are fixed prior to inference and held constant across all proteins unless noted.

Component	Parameter	Value
MSA filtering	Gap fraction threshold	25%
AF-Cluster	Alphabet size	21
	`min_samples`	3
	ϵ scan range	$[3.0, 20.0]$, step 0.5
	ϵ scan subsample	25%
Subsampling	Subsets per arm	12
	Sequences per subset	32
	k-means clusters	12
	Switch columns retained (Arm D)	30
	Arm C weights (α, β, γ)	$(0.5, 0.3, 0.2)$
	Arm C MDS dimensionality	10
	Large-MSA threshold (Arm C)	$N > 3{,}000$
	High-variance percentile	80th
AlphaFold 2	Model type	`alphafold2_ptm`
	Screen: models $\times$ seeds	1×2
	Refine: models $\times$ seeds	5×4
	Recycling iterations	3
	Random seed	0
Refinement	K (refine budget)	$\max(4, \min(16, \lceil 0.2n \rceil))$
	Ambiguity threshold	1.0 Å
	Basin TM-score linkage	0.80
Evaluation	Hit RMSD threshold (common core)	3.0 Å
	Hit pLDDT threshold (overall)	70
	Hit pLDDT threshold (switch region)	70
	Collapse-score TM threshold	0.80
	Common core displacement threshold	1.5 Å
	Switch region displacement threshold	3.0 Å

B.10. Computational Resources

All experiments are conducted on a single server with 8 $\times$ NVIDIA A100-SXM4-80GB GPUs. FrustrAI-Seq inference and ColabFold predictions run on individual GPUs with exclusive allocation per job.

FrustrAI-Seq inference times per case are: KaiB (6,816 sequences, 91 aa): 24 min; Mpt53 (5,638 sequences, 136 aa): 20 min; RfaH (3,350 sequences, 162 aa): 2 min; MAD2 (2,066 sequences, 205 aa): 1 min; GA/GB variants (167-177 sequences, 56 aa): 20-44 s each.

For AlphaFold 2 inference, screen and refine stages for short proteins (56 aa) require approximately 5 and 50 minutes per arm, respectively; for longer proteins (91-136 aa), approximately 10 and 80 minutes per arm. Across all experiments, approximately 17,000 AlphaFold 2 model evaluations were performed in total. Peak GPU memory per job is 12-25 GB for proteins up to 162 residues; AF-Cluster on RfaH exceeded available VRAM due to the large number of DBSCAN clusters (116), and was aborted. This failure is reported in Table 6.