# SMICE: enhanced conformational sampling using AlphaFold and coevolutionary information

**Yongkai Chen** [1] **Samuel W.K. Wong** [2] **S. C. Kou** [1]

## Abstract

Protein structure prediction has been revolutionized by AlphaFold, yet a key limitation remains: its inability to characterize the multiple conformations of fold-switching proteins. Current approaches to address this limitation within the AlphaFold framework rely on subsampling the multiple sequence alignment (MSA) input, either through random sampling or clustering, but these methods are statistically inefficient and fail to utilize coevolutionary information between residues. We introduce SMICE, a sequential sampling framework that systematically explores the MSA space by incorporating residue-specific frequencies and coevolutionary patterns inferred via Markov random fields. On a benchmark set of 92 fold-switching proteins, SMICE outperforms existing methods and substantially improves structural diversity.

## 1. Introduction

It is fair to say that AlphaFold (Jumper et al., 2021) has revolutionized the task of protein structure prediction from sequences. While AlphaFold has received incremental updates since its inception (Baek et al., 2021; Mirdita et al., 2022; Abramson et al., 2024) and continues to be ubiquitous for structure prediction of single protein sequences, it is still challenging to adequately characterize distinct conformational states for a given protein. Proteins are dynamic and change in response to their environment and/or binding partners (Bu & Callaway, 2011); however, each entry in the Protein Data Bank (PDB, Berman et al., 2000) only provides a static experimentally-determined structure for a given sequence. Laboratory techniques to capture protein dynamics are not readily available, and this limitation extends to AlphaFold, which is trained under the one-sequence-to-one-structure paradigm. When predicting the structures of fold-switching proteins, which have at least two distinct yet stable structures, AlphaFold by default can only predict one of the structures (Chakravarty & Porter, 2022).

This limitation has inspired studies into how AlphaFold can be enhanced to provide predictions that capture all possible foldings for a protein sequence. These efforts largely revolve around modifying AlphaFold inputs, most notably in the form of the multiple sequence alignment (MSA). A MSA consists of a list of protein sequences that are evolutionarily or structurally related to the target sequence, usually retrieved by querying large databases of known protein sequences (Johnson et al., 2010; Remmert et al., 2012; Steinegger & Söding, 2017). With a well-constructed MSA, AlphaFold is generally regarded to be more accurate than protein language model-based folding algorithms that operate on the input sequence only, e.g., the ESM family (Hayes et al., 2025) and OmegaFold (Wu et al., 2022). Existing approaches to diversify AlphaFold's predictions have focused on subsampling the full MSA (or generating multiple shallow MSAs) to generate structures from a larger conformational space. These include *random sampling* (Del Alamo et al., 2022; Monteiro da Silva et al., 2024), which draw fixed-size random subsets from the full MSA, and *clustering* (Wayment-Steele et al., 2024), where similar sequences are grouped into clusters that are used as input MSAs.

However, these approaches have two main weaknesses. First, from a statistical perspective, they fail to explore the space of MSA subsets in an effective or unbiased manner. The ideal case for fully leveraging prediction diversity via MSA subsampling would be to exhaustively enumerate all possible subsets, ranging from subsets consisting of diverse sequences to subsets with highly homogeneous sequences, along with intermediate combinations. Clustering methods are inherently biased as they only sample those subsets with highly homogeneous sequences. In contrast, random sampling, though theoretically unbiased, suffers from low sampling efficiency and fails to use any sequence information within the MSA. Second, existing approaches treat each residue in the sequence independently: they compare the amino acid types for each residue (and disregard the inter-

[1]Department of Statistics, Harvard University, Cambridge, MA, United States [2]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada. Correspondence to: S. C. Kou <kou@stat.harvard.edu>, Samuel W.K. Wong <samuel.wong@uwaterloo.ca>.

*Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. 

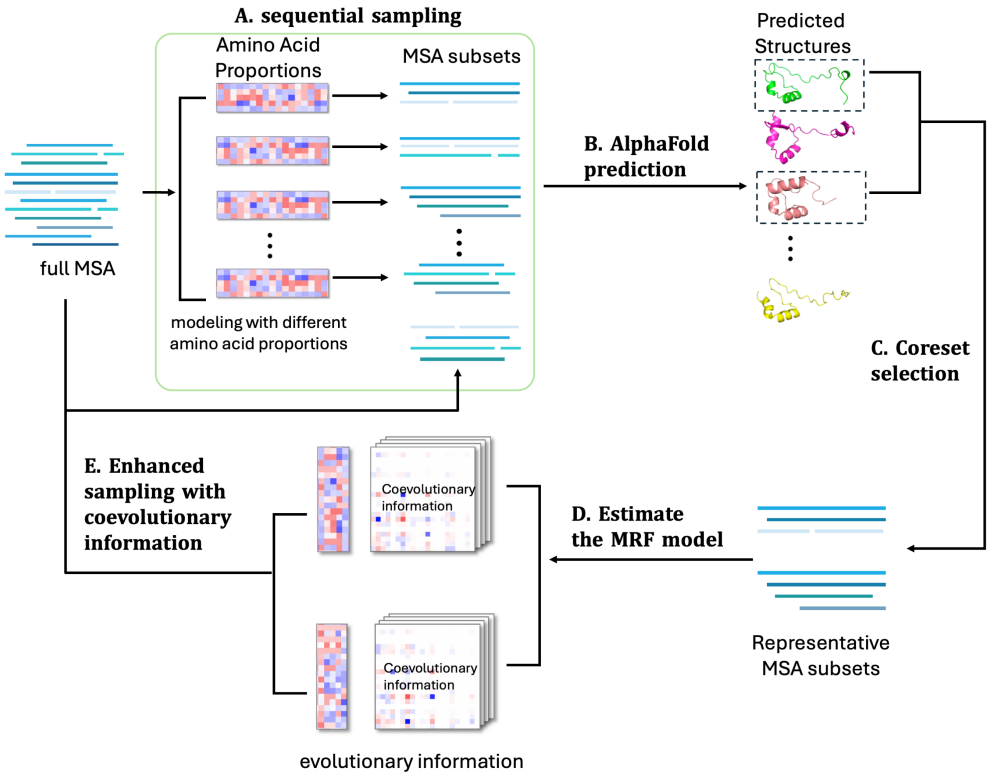


*Figure 1.* Flowchart of SMICE. (A) MSA subsets are drawn from the full MSA using sequential sampling, which are initialized by modeling MSA subsets with different amino acid proportions. (B) Structure predictions are made on the MSA subsets with AlphaFold. (C) Representative predicted structures and their corresponding MSA subsets are extracted via coreset selection on the structures' contact maps. (D) For each representative MSA subset, we estimate its coevolutionary information using a Markov random field (MRF) model. (E) Additional MSA subsets are generated via enhanced sampling, which utilize the differences in coevolutionary information embedded within the representative MSA subsets.

actions between different residues) to guide the selection of MSA subsets. Coevolutionary information in the MSA, which refers to the statistical dependence between two (potentially far apart) residue positions in the MSA (Morcos et al., 2014), is believed to play an important role in how AlphaFold makes predictions (Roney & Ovchinnikov, 2022). Emerging evidence also suggests that the prediction of multiple conformations through MSA subsampling arises from selecting subsets that contain distinct coevolutionary information patterns (Wayment-Steele et al., 2024). However, subsampling strategies based solely on marginal statistics (e.g., amino acid frequencies of each residue) or sequence similarity may fail to capture the dependence between different residues or generate MSA subsets with diverse coevolutionary information.

To address the aforementioned limitations, we propose an iterative sampling method SMICE (**S**ampling **M**SA **I**teratively with **C**o**E**volution information), which formally embeds MSA subsampling into generative probabilistic models. It is constructed as a sequential sampling procedure that incorporates the coevolutionary information into the sampling criterion, as shown in Figure 1. To begin, we sequentially sample sequences based on each residue's marginal frequencies using varying hyperparameter configurations, which ensure high diversity across all sampled subsets. Subsequently, we use the sampled MSA subsets to generate an initial set of structure predictions using AlphaFold. To further increase the diversity of MSA subsets' coevolutionary information, we extract the structurally distinct predictions and analyze the coevolutionary information from each corresponding MSA subset using a Markov random field (MRF) model (Kamisetty et al., 2013). This approach efficiently captures variations in coevolutionary information that contribute to the structural diversity of the predictions. The fitted MRF models then guide the generation of new MSA subsets with increasingly diverse coevolutionary information. Subsequently, we use AlphaFold to predict structures for all sampled MSA subsets. This process is iterated to generate the final structure predictions.

We demonstrate the performance of SMICE on a benchmark set of 92 fold-switching proteins from Chakravarty et al. (2024), where each protein has two distinct conformations. First, we identified several cases where SMICE successfully predicted both conformations, while clustering and random sampling failed to capture at least one conformation. Second, in the analysis of all the fold-switching proteins, SMICE consistently outperforms other methods in predicting both conformations while maintaining high prediction fidelity across most cases. Our method substantially expands the capabilities of MSA subsampling and provides a reliable solution for predicting multiple conformations for the fold-switching proteins.

## 2. Related Research

Several reports of successful MSA subsampling for predicting protein conformations (da Silva et al., 2023; Herrington et al., 2023) have motivated extensive research into its underlying mechanisms and limitations (Chakravarty et al., 2024; Bryant & Noé, 2024; Schafer et al., 2025). Schafer & Porter (2023) found that MSAs of the fold-switching proteins contain different evolutionary information for both foldings. This observation aligns with findings that MSA subsets making distinct predictions are found to contain substantial differences in both marginal amino acid proportions at individual residues and coevolutionary information between residue pairs (Wayment-Steele et al., 2024). However, it remains unclear whether the success arises from capturing different evolutionary information in the MSA subsets or simply from memorizing different sequences within the MSA subsets (Chakravarty et al., 2024). Given the black-box nature of AlphaFold, which makes it difficult to directly examine these factors, a cautious approach is to generate MSA subsets that are as diverse as possible in terms of evolutionary information, thereby increasing prediction diversity and minimizing the risk of overlooking potential causes behind the success. To the best of our knowledge, no prior work has explicitly incorporated coevolutionary information in MSA subset generation.

Besides MSA ablation via subsampling the MSA, alternative approaches have been proposed. Stein & Mchaourab (2022) applies point mutations to the MSA on the predicted contact points. Jing et al. (2024) proposes training a flow-matching variant of AlphaFold, which achieves higher diversity in its predictions compared with simple MSA subsampling.

## 3. Method

As shown in Figure 1, the pipeline of our method consists of the following steps. First, we create MSA subsets via a sequential sampling strategy. Then, we select the representative MSA subsets among all MSA subsets via coreset selection based on their predicted structures. After this, we estimate the coevolutionary information for each representative MSA subset to guide the generation of new subsets. This process is iterated to generate the final structure predictions. We present the details of each step in the following.

### 3.1. Sequential Sampling on the Full MSA

In this section, we develop a sequential sampling method to create MSA subsets with diverse marginal statistics, i.e., the amino acid frequencies of each residue. Let $\mathcal{M} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ denote the set of sequences in the full MSA for the target sequence, where $\mathbf{Y}_i$, a $L \times 22$ binary matrix, is the one-hot encoding of the $i^{th}$ sequence. The 22 categories include the 20 standard amino acids, one for unknown amino acid types, and one for gaps in the alignment. Suppose we start with a small MSA subset $\mathcal{A} \subset \mathcal{M}$. To preserve consistent evolutionary information, our sequential sampling strategy randomly draws a new (i.e., unsampled in the current MSA subset) sequence $\tilde{\mathbf{Y}}$ with an acceptance probability proportional to its similarity to the current MSA subset $\mathcal{A}$. To achieve this, we design the sampling probability of $\tilde{\mathbf{Y}}$ based on how its inclusion changes the amino acid proportions at all residues of $\mathcal{A}$.

Let $\boldsymbol{p}_l$ be the length-22 vector of probabilities for the different amino acid types at the $l^{th}$ residue position, $l = 1, \ldots, L$. For an initialization of $\mathcal{A}$ starting from $\emptyset$, we estimate $\{\boldsymbol{p}_l\}_{l=1}^L$ using a Bayesian approach with a prior distribution specified by a $L \times 22$ matrix $\mathbf{\Pi}$, where the $l^{th}$ row of $\mathbf{\Pi}$ is the prior vector for the $l^{th}$ residue position. Specifically, we model that $\boldsymbol{p}_l$ independently follows a Dirichlet distribution,

$$\boldsymbol{p}_l \sim \text{Dirichlet}(\tau \mathbf{\Pi}^{\mathrm{T}} \boldsymbol{e}_l), \tag{1}$$

where $\tau > 0$ controls the strength of the prior and $\boldsymbol{e}_l$ is the $l^{th}$ unit vector so that $\mathbf{\Pi}^{\mathrm{T}} \boldsymbol{e}_l$ gives the $l^{th}$ row of $\mathbf{\Pi}$.

Given the current MSA subset $\mathcal{A}$ and the prior distribution of $\{\boldsymbol{p}_l\}_{l=1}^L$ in Eq.(1), the maximum a posteriori (MAP) estimate of $\boldsymbol{p}_l$ is

$$\widehat{\boldsymbol{p}}_l(\mathcal{A}, \mathbf{\Pi}) = \frac{(\sum_{\mathbf{Y} \in \mathcal{A}} \mathbf{Y} + \tau \mathbf{\Pi})^{\mathrm{T}} \boldsymbol{e}_l}{|\mathcal{A}| + \tau}. \tag{2}$$

Notice that in the early stages of sequential sampling, the MAP estimate is dominated by $\mathbf{\Pi}$. By assigning high sampling probability to sequences that make small changes to the MAP estimates, sequences that closely match $\mathbf{\Pi}$ will be more favored. This allows us to generate MSA subsets covering different local regions in the sequence space by varying the choices of $\mathbf{\Pi}$. The details of setting $\tau$ and $\mathbf{\Pi}$ are provided in Appendix B.

To decide whether to accept a candidate sequence $\tilde{\mathbf{Y}}$, we first calculate the $L_1$ norm of the MAP estimate's change

by including $\tilde{\mathbf{Y}}$ vs. not including it,

$$\begin{aligned}\Delta_l(\tilde{\mathbf{Y}}, \mathcal{A}, \mathbf{\Pi}) &:= |\widehat{\boldsymbol{p}}_l(\mathcal{A}, \mathbf{\Pi}) - \widehat{\boldsymbol{p}}_l(\mathcal{A} \cup \{\tilde{\mathbf{Y}}\}, \mathbf{\Pi})|_1 \\ &= \frac{|\widehat{\boldsymbol{p}}_l(\mathcal{A}, \mathbf{\Pi}) - \tilde{\mathbf{Y}}^{\mathrm{T}} \boldsymbol{e}_l|_1}{|\mathcal{A}| + \tau + 1}. \end{aligned} \tag{3}$$

To measure the overall change of the MAP estimate of $\{\boldsymbol{p}_l\}_{l=1}^L$, we consider two sources of heterogeneity across all residues: (1) different residues have different levels of evolutionary conservation, and (2) residues with many gaps may provide less reliable information. To address them, we rescale the changes for each residue through

$$Q_l(\tilde{\mathbf{Y}}, \mathcal{A}, \mathbf{\Pi}) = w_l \frac{\Delta_l(\tilde{\mathbf{Y}}, \mathcal{A}, \mathbf{\Pi})}{\sqrt{Var(\Delta_l(\tilde{\mathbf{Y}}, \mathcal{A}, \mathbf{\Pi}))}}, \tag{4}$$

where $Var(\Delta_l(\tilde{\mathbf{Y}}, \mathcal{A}, \mathbf{\Pi}))$ is computed by assuming the marginal distribution of the $l^{th}$ row of $\tilde{\mathbf{Y}}$ is the one-trial multinomial distribution with probability vector $\widehat{\boldsymbol{p}}_l(\mathcal{A}, \mathbf{\Pi})$, and $w_l$ is the proportion of gaps in the $l^{th}$ position for the full MSA. Appendix A provides the details of Eq.(4).

We now calculate the acceptance probability of the randomly drawn candidate sequence $\tilde{\mathbf{Y}}$ given the current MSA subset $\mathcal{A}$ through

$$\min\left\{1, \exp\{-\frac{\lambda}{L}\sum_{l=1}^{L} Q_l(\tilde{\mathbf{Y}}, \mathcal{A}, \mathbf{\Pi})\}/C\right\}, \tag{5}$$

where $C$ is a tuning parameter controlling the overall sample size, and $\lambda \geq 0$ controls the homogeneity level in the sampled MSA subsets. When $\lambda$ is larger, only sequences resulting in small changes to the MAP estimate are likely to be accepted. This leads to highly homogeneous subsets where sampled sequences are tightly clustered. In contrast, a small $\lambda$ allows for greater diversity in the sampled MSA subsets. In the limiting case of $\lambda = 0$, the sequential sampling strategy reduces to uniform random sampling. For broader exploration, we recommend sampling with multiple $\lambda$ values (see Appendix B).

After we query all the sequences, or the size of the MSA subset $\mathcal{A}$ reaches the preset maximum size, we use $\mathcal{A}$ as the input MSA for AlphaFold. By repeating this procedure multiple times with varying choices of the hyperparameters $\mathbf{\Pi}$ and $\lambda$, we induce a set of MSA subsets and its corresponding AlphaFold predicted conformations.

### 3.2. Selection of Representative MSA Subsets

From the sequentially sampled MSA subsets, we now use coreset selection to select the most representative subsets based on their predicted structures. Given the predicted conformations obtained in the previous step, we expect that the distinct structures predicted also exhibit distinct evolutionary information in their corresponding MSA subsets. Thus, extracting representative structures and their MSA subsets could provide valuable information for generating new MSA subsets that lead to more diverse structures.

Specifically, we first straighten the $L \times L$ contact map matrices of all the predicted structures into vectors, and principal component analysis (PCA) is applied to obtain low-dimensional representations of the contact map. The number of PCs is selected to explain at least 90% of the total variance. We then apply coreset selection (Feldman, 2020) on the PCA coordinates to extract $K$ distinct structures[1] (see Appendix B for the details of the coreset selection). Finally, for each of the $K$ structures, we identify its corresponding MSA subset; these $K$ MSA subsets will be used to extract the coevolutionary information.

### 3.3. Enhanced Sampling using Coevolutionary Information

By analyzing differences in coevolutionary information across representative MSA subsets, we develop an enhanced sampling strategy to generate more diverse samples. A commonly used statistical model for analyzing the coevolutionary information of the MSA is the Markov random field (MRF) model (Kamisetty et al., 2013) with the parameter $\mathbf{\Theta} = \left\{\{\mathbf{V}_l\}_{l=1}^L, \{\mathbf{W}_{l,m}\}_{1\leq l<m\leq L}^L\right\}$,

$$\mathrm{P}_{MRF}(\mathbf{Y}|\mathbf{\Theta}) \propto \exp\left(\sum_{l=1}^{L} \boldsymbol{e}_l^{\mathrm{T}}\mathbf{Y}\left[\mathbf{V}_l + \sum_{m>l}^{L} \mathbf{W}_{l,m}\mathbf{Y}^{\mathrm{T}}\boldsymbol{e}_m\right]\right), \tag{6}$$

where $\mathbf{V}_l$ is a 22-dimensional vector capturing the marginal effect of the amino acid type at the $l^{th}$ residue and $\mathbf{W}_{l,m}$ is a $22 \times 22$ matrix capturing the pairwise interaction effect between the $m^{th}$ and $l^{th}$ residues.

Sampling MSA subsets with diverse coevolutionary information can enable AlphaFold to more effectively explore the conformational space. To leverage this coevolutionary information efficiently, we first estimate the MRF model for each of the representative MSA subsets extracted in the previous steps using GREMLIN (Kamisetty et al., 2013), see Appendix B for more details. Suppose there are $K$ MSA subsets, so that the MRF model is estimated $K$ times. Then, for each pair of estimated MRF models, denoting the parameters by $\mathbf{\Theta}$ and $\mathbf{\Theta}'$, we calculate the probability ratio $\mathrm{P}_{MRF}(\mathbf{Y}_i|\mathbf{\Theta})/\mathrm{P}_{MRF}(\mathbf{Y}_i|\mathbf{\Theta}')$ across all sequences in the full MSA. A high ratio indicates that a sequence is more consistent with the coevolutionary information captured by

[1]Note that AlphaFold is an ensemble model with five different sets of model weights, i.e., each MSA input generates five different structure predictions by AlphaFold. Therefore, the extraction of representative MSA subsets and all subsequent steps are performed independently for each of the five weight sets of AlphaFold.

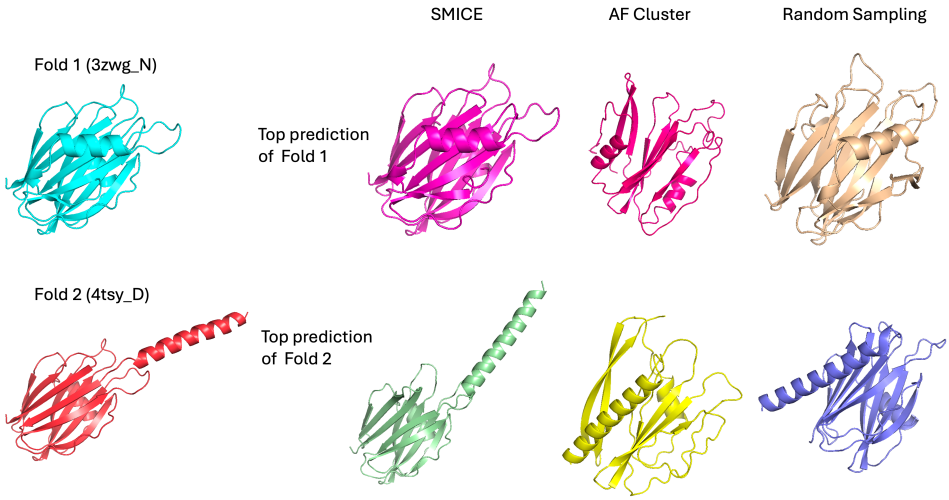


*Figure 2.* Visualization of a fold-switching protein with two conformations (3zwgN and 4tsyD) and the top predictions for each conformation by the three methods.

$\boldsymbol{\Theta}$ compared to $\boldsymbol{\Theta}'$. By pooling sequences with high probability ratios into a new MSA subset, we may strengthen their shared coevolutionary information. Finally, for each pair of estimated MRF models, we rank all sequences by their computed probability ratios and select the top $N$ sequences, where $N$ is a predetermined subset size. These $N$ sequences are used as one input MSA subset for AlphaFold to generate new predictions (see Appendix B for the details of setting $N$). Since there are $K(K-1)$ such ordered pairs, this procedure generates $K(K-1)$ MSA subsets.

The three steps (detailed in Sections 3.1 to 3.3) are iterated in SMICE a few times (usually two or three rounds) to obtain the final structure predictions.

## 4. Experiments on the fold-switching proteins

We examine the performance of our methodology on a set of 92 fold-switching proteins in Chakravarty et al. (2024). Each protein has two conformations, and the residues corresponding to the fold-switching region (fsr) have been previously identified. To assess prediction accuracy, we computed TMscore, a widely-used metric for evaluating the structural similarity between two structures, for a combined region that includes the fsr and its neighboring segment of half the fsr's length[2]. The cases where the full MSA contains fewer than 20 sequences are removed. We compared SMICE with AF-Cluster and random sampling. The implementation details are provided in Appendix C.

[2]As noted by Chakravarty et al. (2024), whole-protein TMscores tend to overestimate prediction accuracy for fsr's. Moreover, we found that TMscores computed solely on the fsr often fail to distinguish conformations resulting from hinge motions (Bryant & Noé, 2024), where the global fold remains largely unchanged but relative orientations between regions shift.

### 4.1. Illustrative Examples

To highlight the advances of SMICE, we begin by presenting the results for three illustrative fold-switching proteins, each with two distinct conformations (PDB IDs: 3zwgN/4tsyD, 2c1vB/2c1uC, and 3hdeA/3hdefA).

Results are shown in Figure 3, with each row corresponding to one fold-switching protein. The left panel in each row presents the results from SMICE, the middle panel shows the results from AF-Cluster, and the right panel shows the results from random sampling. In each scatter plot, the TMscores of the predicted structures to the two known conformations are plotted on the $x$ and $y$ axes, respectively. Since AF-Cluster generates a variable number of predictions (depending on the number of clusters), to ensure a fair comparison, we resample the same number of predictions as AF-Cluster from the prediction sets of our proposed method. Similarly, we repeat random sampling to generate the same number of MSA subsets as AF-Cluster.

We found that MSA subsets generated by SMICE have predictions with higher TMscore for both conformations across all three cases. In contrast, MSA subsets from AF-Cluster failed to make accurate predictions for the alternative conformation 3zwgN (Figure 3A) and 2c1uC (Figure 3B), while random sampling failed for all three cases. Furthermore, AF-Cluster makes less reliable predictions (i.e., in regions of the conformational space not close to either fold) compared to SMICE. For example, AF-Cluster makes significantly larger proportions of predictions with both TMscores lower than 0.6 in case 3 compared to SMICE, as shown in Figure 3(C).

For the fold-switching protein with conformations represented by 3zwgN and 4tsyD, Figure 2 visualizes the structures corresponding to the best predictions generated by the three methods. The top and bottom rows display the

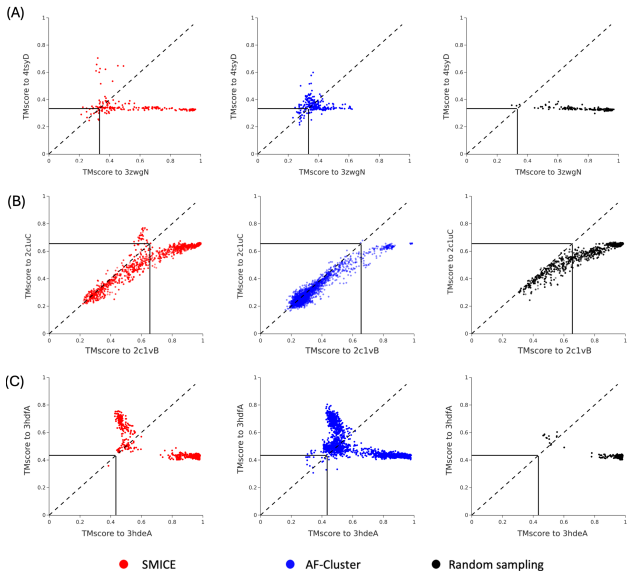


*Figure 3.* TMscores of the AlphaFold prediction on the MSA subsets supplied by SMICE, AF-Cluster, and random sampling on three fold-switching proteins. The results from SMICE are shown in red. The results from MSA subsets obtained via sequence clustering in AF-Cluster are shown in blue, while those from random sampling are shown in black. The black vertical and horizontal lines represent the TMscores between the two conformations of the fold-switching proteins. (A) 3zwgN/4tsyD (B) 2c1vB/2c1uC (C) 3hdeA/3hdefA.

structures with the highest TMscores to Fold1 (3zwgN) and Fold2 (4tsyD), respectively; only SMICE was able to generate a structure that resembles Fold2.

Additional examples are provided in Appendix D.

### 4.2. Comparison with Other Methods

We next provide a comprehensive comparison against AF-cluster and random sampling on the full set of proteins.

First, taking each set of predictions generated for a given fold-switching protein, we check if the structures (1) predict Fold1[3], (2) predict Fold2, or (3) fail to predict either fold. To measure these, we compare the TMscores of both folds against a predefined threshold: a prediction is considered to successfully predict Fold1 if TMscore1 (TMscore to Fold1) exceeds both TMscore2 (TMscore to Fold2) and the threshold, to successfully predict Fold2 if TMscore2 exceeds both TMscore1 and the threshold, and as a failure if neither TMscore1 nor TMscore2 surpasses the threshold. By varying the threshold, we calculate the proportions of predictions of these three cases across different levels of prediction confidence for each competing method. Intuitively, only those

[3]For each pair of conformational structures, we use the Fold1 and Fold2 defined by Chakravarty et al. (2024), where Fold2 is more challenging to predict and usually exhibits lower TMscores compared to Fold1.

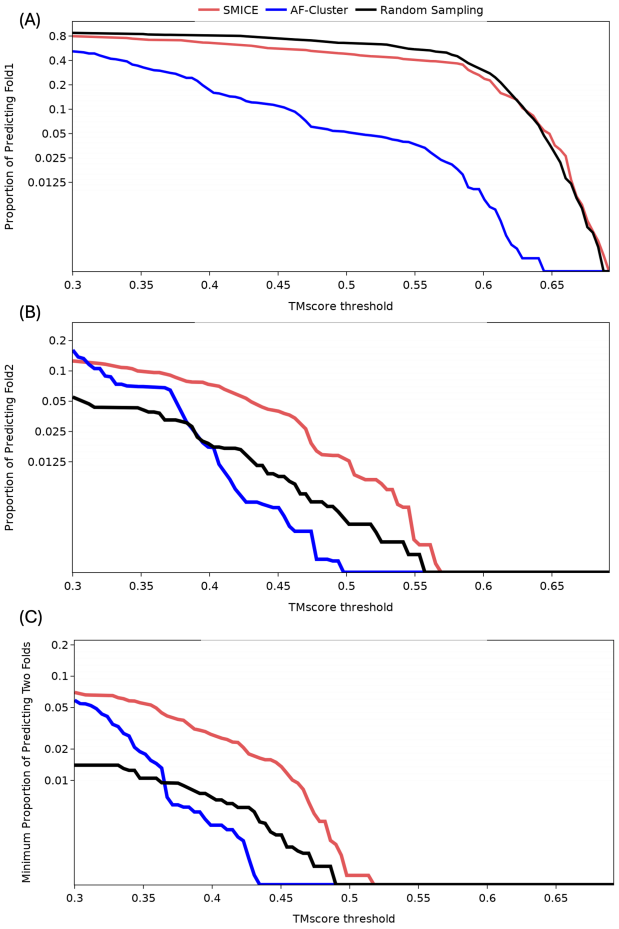


*Figure 4.* Comparison results of SMICE against AF-Cluster and Random Sampling on the proportions of predicting Fold1, the proportion of predicting Fold2, and the minimum proportions of predicting two folds. For varying TM-score thresholds, we compute the median of the proportions across all fold-switching proteins. Different methods are colored differently

prediction sets with high proportions for both predicting Fold1 and Fold2 can be considered as successful. Therefore, we use the minimum of the two prediction proportions as a performance metric, representing the ability to predict both folds simultaneously.

In Figure 4, we summarize the results across all fold-switching proteins with the median of the prediction proportions for Fold1, the prediction proportions for Fold2, and the minimum proportions for predicting both folds. The plotted curves represent the median of the proportions over the 92 fold-switching proteins, as a function of the TMscore threshold. It is seen that SMICE consistently outperforms AF-Cluster and random sampling, achieving the highest prediction proportions for both folds and the highest minimum success rates across nearly all thresholds and quantiles. An interesting finding is that random sampling performs comparably or slightly better than SMICE in predicting Fold1. This may be due to the random sampling's tendency to gener-

ate MSA subsets that share similar evolutionary information with the full MSAs, leading to their predictions being closer to the more easily predictable Fold1 conformation.

We then compare the best prediction result for each method on each fold-switching protein. For each fold-switching protein, we compute three metrics on its prediction sets: (1) max-TMscore1, i.e., the maximum TMscore to Fold1; (2) max-TMscore2, i.e., the maximum TMscore to Fold2; (3) minimax-TMscore, i.e., the minimum of max-TMscore1 and max-TMscore2, reflecting the worst-case prediction gap for either conformation. A high minimax-TMscore means that both folds are accurately predicted.

For a fair comparison, we repeatedly resample the same number of predictions as AF-Cluster from the prediction sets of SMICE and random sampling 500 times, and calculate the averaged metrics across all repetitions.

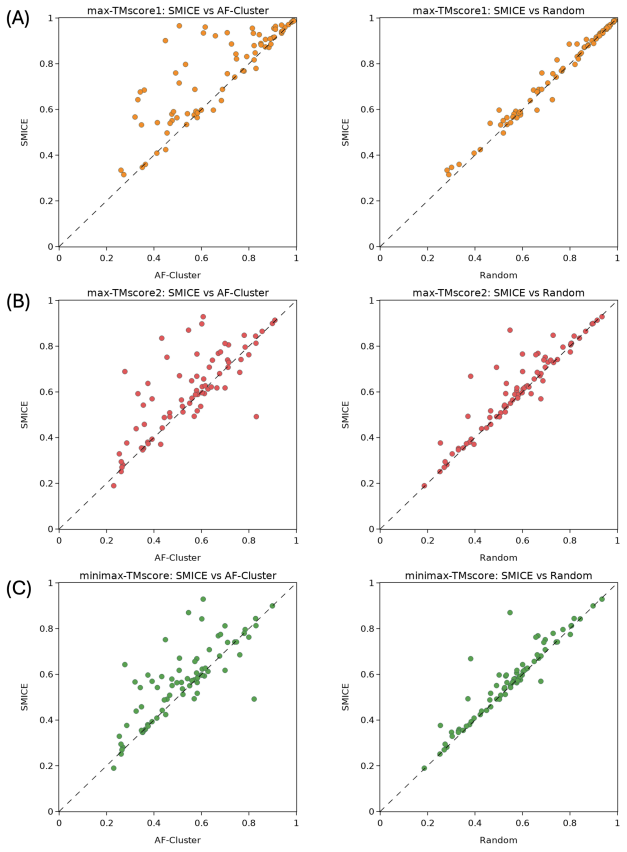


*Figure 5.* Comparing the results of SMICE against AF-Cluster and Random Sampling. Each point represents one fold-switching protein. Points above the $45^\circ$ dashed line indicate cases where SMICE achieves higher scores than the competing methods.

The results comparing SMICE against AF-Cluster and random sampling are shown in Figure 5. The rows display the three evaluation metrics: (A) max-TMscore1, (B) max-TMscore2, (C) minimax-TMscore. In each row, the left panel shows the comparison between SMICE and AF-Cluster, while the right panel shows the comparison with random sampling. We found that SMICE outperforms AF-Cluster in capturing both conformations in most cases, as shown in Figure 5. Random sampling performs comparably to SMICE for predicting Fold1 (structures that are easier to predict using AlphaFold without subsampling), while it is outperformed by SMICE in predicting Fold2. This result is consistent with the finding in Figure 4.

## 5. Conclusion

Our work addresses a critical limitation of AlphaFold in predicting the multiple conformations of fold-switching proteins. We reformulate MSA subsampling as a probabilistic, coevolutionary information-aware procedure. By varying hyperparameters that control sequence homogeneity and initializations, we avoid the redundancy of random sampling and the biases toward sampling highly homogeneous sequences of clustering. We also explicitly account for coevolutionary information by modeling residue dependence, further increasing the diversity of coevolutionary information in the MSA subsets. These properties ensure a significant improvement in the diversity and accuracy of predicted conformations compared to existing methods. While we focus on fold-switching proteins, our framework could be potentially generalized to other problems, e.g., predicting the dynamic conformations of intrinsic dynamic proteins (IDPs).

## Acknowledgements

S.W.K. Wong was partially supported by Discovery Grant RGPIN-2019-04771 from the Natural Sciences and Engineering Research Council of Canada. S.C. Kou acknowledges support from the Harvard Data Science Initiative.

## Impact Statement

This work lies at the intersection of AI, structural biology, and statistics. The proposed method significantly enhances MSA subsampling strategies, enabling AlphaFold to better predict multiple protein conformations. This advancement has broad potential impact on genomics, functional annotation, protein engineering, and biomedical research.

## References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630: 493–500, 2024.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

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

Bu, Z. and Callaway, D. J. Proteins move! Protein dynamics and long-range allostery in cell signaling. *Advances in protein chemistry and structural biology*, 83:163–221, 2011.

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.

da Silva, G. M., Cui, J. Y., Dalgarno, D. C., Lisi, G. P., and Rubenstein, B. M. Predicting relative populations of protein conformations without a physics engine using AlphaFold 2. *bioRxiv*, 2023.

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.

Feldman, D. Core-sets: Updated survey. *Sampling techniques for supervised or unsupervised tasks*, pp. 23–44, 2020.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Herrington, N. B., Stein, D., Li, Y. C., Pandey, G., and Schlessinger, A. Exploring the druggable conformational space of protein kinases using AI-generated structures. *bioRxiv*, pp. 2023–08, 2023.

Jing, B., Berger, B., and Jaakkola, T. AlphaFold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 22277–22303, 2024.

Johnson, L. S., Eddy, S. R., and Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics*, 11:1–8, 2010.

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.

Kamisetty, H., Ovchinnikov, S., and Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39): 15674–15679, 2013.

Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. ColabFold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Monteiro da Silva, G., Cui, J. Y., Dalgarno, D. C., Lisi, G. P., and Rubenstein, B. M. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nature communications*, 15(1):2464, 2024.

Morcos, F., Schafer, N. P., Cheng, R. R., Onuchic, J. N., and Wolynes, P. G. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences*, 111(34):12408–12413, 2014.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 9(2):173–175, 2012.

Roney, J. P. and Ovchinnikov, S. State-of-the-art estimation of protein model accuracy using AlphaFold. *Physical Review Letters*, 129(23):238101, 2022.

Schafer, J. W. and Porter, L. L. Evolutionary selection of proteins with two folds. *Biophysical Journal*, 122(3): 474a, 2023.

Schafer, J. W., Lee, M., Chakravarty, D., Thole, J. F., Chen, E. A., and Porter, L. L. Sequence clustering confounds AlphaFold2. *Nature*, 638(8051):E8–E12, 2025.

Stein, R. A. and Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLOS Computational Biology*, 18 (8):e1010483, 2022.

Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

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.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022.

# A. Mathematical Details

By assuming the marginal distribution of the $l^{th}$ row of $\tilde{\mathbf{Y}}$ is the one-trial multinomial distribution with probability vector as $\widehat{\boldsymbol{p}}_l(\mathcal{A}, \boldsymbol{\Pi})$, we have

$$E[\Delta_l(\tilde{\mathbf{Y}}, \mathcal{A}, \boldsymbol{\Pi})] = \frac{E|\widehat{\boldsymbol{p}}_l(\mathcal{A}, \boldsymbol{\Pi}) - \mathbf{Y}^{\mathrm{T}}\boldsymbol{e}_l|_1}{|\mathcal{A}| + \tau + 1} = 2\sum_{a=1}^{22} \frac{p_{l,a}(1 - p_{l,a})}{|\mathcal{A}| + \tau + 1},$$

$$Var[\Delta_l(\tilde{\mathbf{Y}}, \mathcal{A}, \boldsymbol{\Pi})] = 4\sum_{a=1}^{22} \frac{p_{l,a}(1 - p_{l,a})^2}{(|\mathcal{A}| + \tau + 1)^2} - E^2[\Delta_l(\tilde{\mathbf{Y}}, \mathcal{A}, \boldsymbol{\Pi})],$$

where $p_{l,a}$ is the $a^{th}$ entry of $\widehat{\boldsymbol{p}}_l(\mathcal{A}, \boldsymbol{\Pi})$.

# B. Additional Details of the Methods

## B.1. Setting Hyperparameters $\tau$ and $\boldsymbol{\Pi}$

As mentioned in Section 3, to promote diversity among the sampled MSA subsets, we vary the choices of $\boldsymbol{\Pi}$ in the prior distribution of the expected amino acid proportions.

We set these values in a data-driven manner. Specifically, for each sequence $\mathbf{Y}_i \in \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$, let $\mathcal{S}(i)$ represent the indices of $\mathbf{Y}_i$'s N-nearest sequences among the full MSA using the Hamming distance. We compute the average proportions of amino acids in all residues across $\mathbf{Y}_i$'s neighbors, $\bar{\mathbf{Y}}_i = \frac{1}{N}\sum_{j \in \mathcal{S}(i)} \mathbf{Y}_j$. We then apply K-medoids clustering to these average amino acid proportions $\{\bar{\mathbf{Y}}_i\}_{i=1}^n$ and the resulting cluster centers. These cluster centers are used as K different choices for $\boldsymbol{\Pi}$. In our experiment, we set $K = 10$. For MSAs with fewer than 100 sequences, we set $N = 10$, and for those with more than 100 sequences, we conduct the above procedure for both $N = 10$ and 30. Finally, we set the strength of the prior $\tau$ as $0.5N$.

## B.2. Setting Hyperparameters in the Acceptance Probabilities

Since $\lambda$ controls the homogeneity of the sequences within the sampled MSA subset, using multiple choices of $\lambda$ could also lead to diverse MSA subsets in terms of evolutionary information. In particular, we choose $\lambda$ from $\{0, 1, 2, 3\}$. Moreover, $C$ in Eq.(5) controls the overall acceptance probability. The common choice of $C$ is set as $\text{argmax}_{\mathbf{Y}} \exp\{-\frac{\lambda}{L}\sum_{l=1}^{L} Q_l(\tilde{\mathbf{Y}}, \mathcal{A}, \boldsymbol{\Pi})\}$ so that the acceptance probability can be no greater than 1.

## B.3. Sequential Sampling Algorithm

The details of the sequential sampling algorithm are summarized in Algorithm 1. Notice that we set a maximum subset size $M$ for efficient computation when the full MSA size is larger. The empirical suggestion of $M$ is as follows: $M = 20$ if the full MSA size $< 100$, $M = 100$ if the full MSA size $\in [100, 500]$, and $M = 200$ otherwise.

**Algorithm 1** Sequential Sampling for MSA Subsets

**Input:** Full MSA $\mathcal{M}$, target maximum subset size $M$, $\boldsymbol{\Pi}$'s configuration set $\{\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_K\}$.
**for** each $\lambda \in \{0, 1, 2, 3\}$ **do**
  **for** each $\boldsymbol{\Pi} \in \{\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_K\}$ **do**
    Initialize $\mathcal{A} \leftarrow \emptyset$
    **while** $|\mathcal{A}| < M$ **do**
      Sample candidate $\tilde{\mathbf{Y}}$: $\tilde{\mathbf{Y}} \leftarrow$ sample from $\mathcal{M}$ without replacement according to the probability in Eq.(5),
      $\mathcal{A} \leftarrow \mathcal{A} \cup \{\tilde{\mathbf{Y}}\}$
    **end while**
    Store $\mathcal{A}$ in output collection if $|\mathcal{A}| > 10$
  **end for**
**end for**
**Output:** All sampled MSA subsets $\mathcal{A}$

### B.4. Coreset Selection for Extracting Representative Structures

The details of the coreset selection are summarized in Algorithm 2.

**Algorithm 2** Coreset Selection

**Input:** The PCA coordinates of the predicted structures' contact maps $\{\mathbf{C}_1, \ldots, \mathbf{C}_N\}$, target coreset size $K$.
Initialize coreset $\mathcal{C} = \{\text{argmax}_{\mathbf{C}\in\{\mathbf{C}_1,\ldots,\mathbf{C}_N\}} \|\mathbf{C}\|_2\}$
**while** $|\mathcal{C}| < K$ **do**
  $\tilde{\mathbf{C}} \leftarrow \text{argmax}_{\mathbf{C}\in\{\mathbf{C}_1,\ldots,\mathbf{C}_N\}} \min_{\mathbf{C}'\in\mathcal{C}} \|\mathbf{C} - \mathbf{C}'\|_2$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{\tilde{\mathbf{C}}\}$
**end while**
**Output:** The selected coreset $\mathcal{C}$ and their corresponding MSA subsets.

### B.5. Estimating the MRF Model with GREMLIN

Given the MRF model in Eq.(6) and sequences $\{\tilde{\mathbf{Y}}_i\}_{i=1}^n$, the log-likelihood of $\boldsymbol{\Theta}$ is

$$l(\boldsymbol{\Theta}|\{\tilde{\mathbf{Y}}_i\}_{i=1}^n) = \sum_{i=1}^{n}\sum_{l=1}^{L} \boldsymbol{e}_l^{\mathrm{T}}\tilde{\mathbf{Y}}_i \left[\mathbf{V}_l + \sum_{m>l}^{L} \mathbf{W}_{l,m}\tilde{\mathbf{Y}}_i^{\mathrm{T}}\boldsymbol{e}_m\right] - n\log\left\{\sum_{\mathbf{Y}} \exp\left(\sum_{l=1}^{L} \boldsymbol{e}_l^{\mathrm{T}}\mathbf{Y}\left[\mathbf{V}_l + \sum_{m>l}^{L}\mathbf{W}_{l,m}\mathbf{Y}^{\mathrm{T}}\boldsymbol{e}_m\right]\right)\right\}, \tag{7}$$

which is highly challenging to maximize directly due to the large number of values of $\mathbf{Y}$ in the summation. Given the high-dimensionality of $\mathbf{V}_l$ and $\mathbf{W}_{m,l}$, the GREMLIN method considers the penalized log pseudo-likelihood,

$$ppl(\boldsymbol{\Theta}|\{\tilde{\mathbf{Y}}_i\}_{i=1}^n) = \sum_{i=1}^{n}\sum_{l=1}^{L} \log P(\tilde{\mathbf{Y}}_{i,l}|\tilde{\mathbf{Y}}_{i,-l}, \boldsymbol{\Theta}) - \lambda_1 \sum_{l=1}^{L}\|\mathbf{V}_l\|_2^2 - \lambda_2 \sum_{m=1}^{L}\sum_{l=1}^{L}\|\mathbf{W}_{l,m}\|_F^2, \tag{8}$$

where $\tilde{\mathbf{Y}}_{i,l}$ is the one-hot encoding of the amino acid at position $l$ in the $i^{th}$ sequence, and $\tilde{\mathbf{Y}}_{i,-l}$ is the one-hot encoding of the amino acid at all other positions in the $i^{th}$ sequence,

$$\log P(\tilde{\mathbf{Y}}_{i,l}|\tilde{\mathbf{Y}}_{i,-l}, \boldsymbol{\Theta}) = \tilde{\mathbf{Y}}_{i,l}^{\mathrm{T}}\mathbf{V}_l + \sum_{m=1,m\neq l}^{L} \tilde{\mathbf{Y}}_{i,l}^{\mathrm{T}}\mathbf{W}_{l,m}\tilde{\mathbf{Y}}_{i,m} - \log\sum_{\boldsymbol{y}_l} \exp\left[\tilde{\mathbf{Y}}_{i,l}^{\mathrm{T}}\mathbf{V}_l + \sum_{m=1,m\neq l}^{L} \boldsymbol{y}_l^{\mathrm{T}}\mathbf{W}_{l,m}\boldsymbol{y}_m\right],$$

where $\boldsymbol{y}_l$ is the one-hot encoding of the amino acid type at position $l$. The GREMLIN method then estimates $\boldsymbol{\Theta}$ by maximizing Eq.(8) with gradient descent using Adam (Kingma, 2014).

### B.6. Enhanced Sampling using Coevolutionary Information

The details of enhanced sampling using coevolutionary information are summarized in Algorithm 3.

## C. Experiment Implementation Details

### C.1. MSA Generation

MSAs were generated using MMseqs2 (Steinegger & Söding, 2017) implemented in ColabFold (Mirdita et al., 2022) by querying the UniRef30 database (Suzek et al., 2015) of known sequences. Sequences are filtered to retain only those with $\geq 90\%$ identity to the target sequence. The minimum coverage required for the MSA sequences is $75\%$. The MSA is enforced to have at most 4,096 sequences.

### C.2. SMICE, AF-Cluster and Random Sampling

As discussed in Appendix B, we consider 80 different hyperparameter configurations for MSA when the number of sequences exceeds 100, and 40 configurations when it is below 100. These configurations consist of four levels of $\lambda$, and 20 choices of $\boldsymbol{\Pi}$ for deep MSAs (sequence number $\geq 100$) or 10 choices for shallow MSAs (sequence number $< 100$). For each

**Algorithm 3** Enhanced Sampling using Coevolutionary Information

**Input:** MSA subsets from the sequential sampling $\mathcal{A}_1, \ldots, \mathcal{A}_N$, target iteration times $n_{iter}$, target coreset size $K$, target MSA subset size in the enhanced sampling $M$.
Make predictions on the MSA subsets $\mathcal{A}_1, \ldots, \mathcal{A}_N$ across all five AlphaFold models
**for** each AlphaFold model **do**
  Obtain the PCA coordinates of the predicted structures' contact maps $\{\mathbf{C}_1, \ldots, \mathbf{C}_N\}$ for the AlphaFold model
  **while** iteration number does not exceed $n_{iter}$ **do**
    Apply Algorithm 2 to $\{\mathbf{C}_1, \ldots, \mathbf{C}_N\}$, and extract $K$ representative MSA subsets $\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_K$
    **for** k = 1,...,K **do**
      Estimate the MRF model's parameter $\mathbf{\Theta}_k$ for the MSA subset $\tilde{\mathcal{A}}_k$
    **end for**
    **for** each pair of $(j, l)$ with $1 \leq j, l \leq K$, $j \neq l$ **do**
      obtain the MSA subset $\breve{\mathcal{A}}_{j,l}$ with size $M$ such that every sequence $\breve{\mathbf{Y}} \in \breve{\mathcal{A}}_{j,l}$ has a larger value of $\frac{\mathrm{P}_{MRF}(\breve{\mathbf{Y}}|\mathbf{\Theta}_j)}{\mathrm{P}_{MRF}(\breve{\mathbf{Y}}|\mathbf{\Theta}_j)}$ than other sequences
    **end for**
    store all $\breve{\mathcal{A}}_{j,l}$ in the set of MSA subsets
    make predictions on the MSA subsets with AlphaFold
    Obtain the PCA coordinates of the new predicted structures' contact maps $\{\mathbf{C}_1, \ldots, \mathbf{C}_N\}$ for the AlphaFold model, where $N = K(K-1)$
  **end while**
**end for**
**Output:** The selected coreset $\mathcal{C}$ and their corresponding MSA subsets.

configuration, we randomly draw four subsets using sequential sampling. We exclude any MSA subsets containing fewer than 10 sequences to ensure reliable AlphaFold predictions. Each remaining MSA subset is used to generate structures using ColabFold v1.5 with default settings, which outputs five predictions per run using AlphaFold2. No template structure is used in the prediction. The enhanced sampling procedure is iteratively repeated twice ($n_{iter} = 2$), and we set the target coreset size $K$ in Algorithm 2 as five, the target MSA subset size in the enhanced sampling as $M = 20$ or $= 100$ if the full MSA has more than 100 sequences. This results in 200 predictions for deep MSA cases and 100 predictions for shallow MSA cases per iteration. In total, we obtain up to 1,800 predictions for deep MSAs and up to 900 predictions for shallow MSAs.

The results of AF-Cluster and random sampling are produced using the default setup of the AF-Cluster pipeline (Wayment-Steele et al., 2024). In the pipeline, uniform random sampling without placement is run up to 200 times on each MSA with sample sizes $= 10$ and 100 (if the full MSA has more than 100 sequences). The predictions are made using ColabFold v1.5 with default settings.

## D. Additional Experiment Results

Figure 6 presents more results demonstrating the diversity and effectiveness of SMICE in conformational sampling compared with AF-Cluster and random sampling.

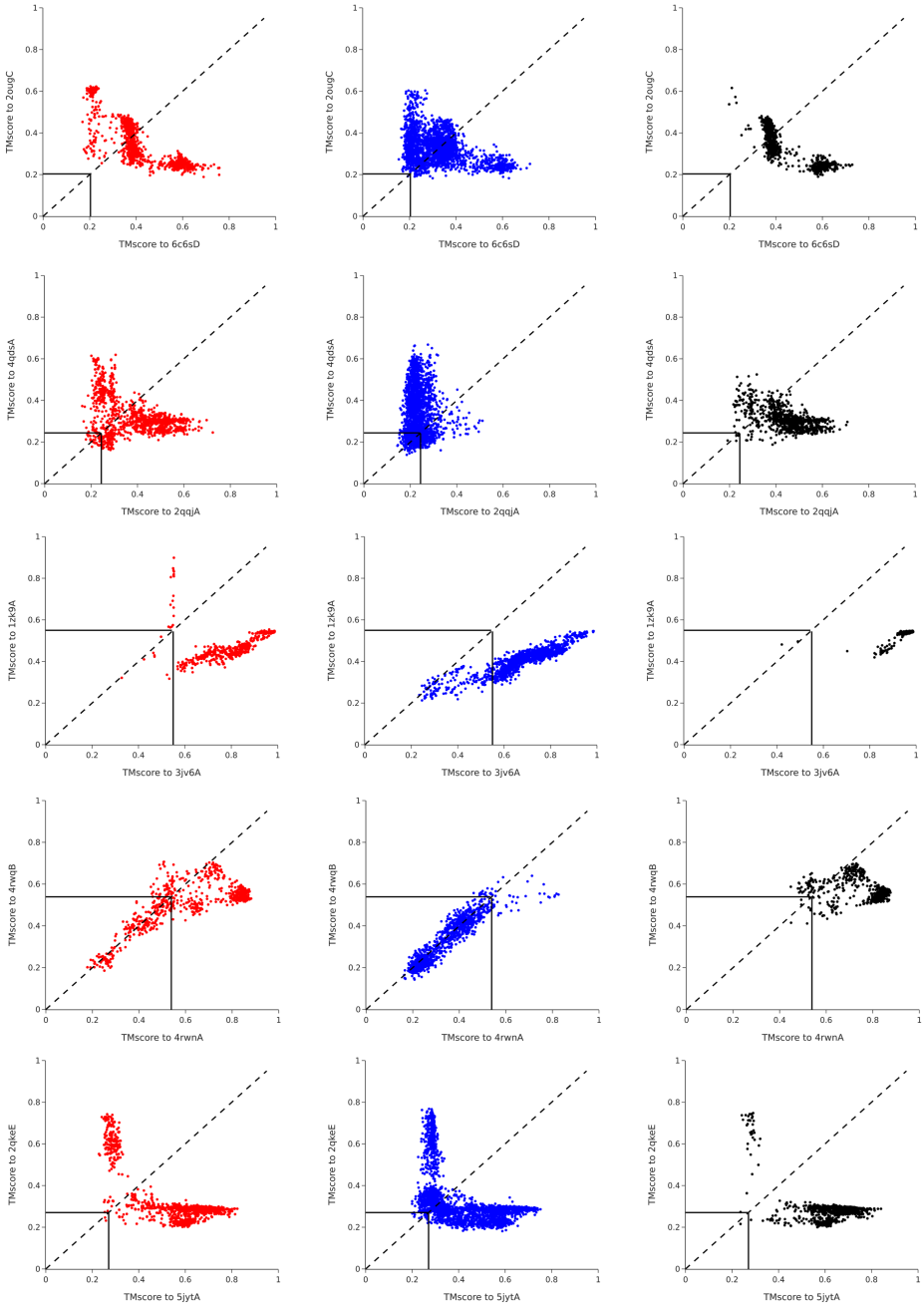


*Figure 6.* Additional examples of fold-switching proteins.