JCTC Journal of Chemical Theory and Computation

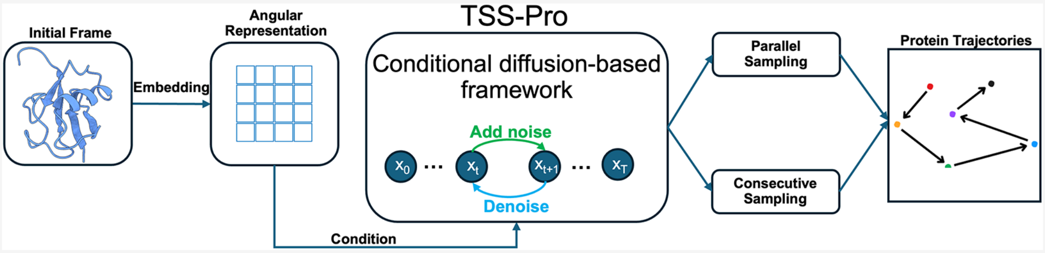
pubs.acs.org/JCTC Article

Efficient Sampling of Short Protein Trajectories with Conditional Diffusion Models

Chuanye Xiong, Palanisamy Kandhan, Dongyang Chen, Zerui Ma, Eleanor D. Smith, and Peng Tao*

Cite This: *J. Chem. Theory Comput.* 2026, 22, 78–94 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: Understanding how protein structures dictate their diverse biological functions remains one of the central and enduring challenges in structural biology. The development of AlphaFold and ESMAtlas marks a significant advance in protein science, enabling the reliable prediction of protein structure directly from amino acid sequence. This advance in structure prediction underscores the need for complementary methods that can explore conformational space and enable efficient sampling of dynamic trajectories. Here, we present TSS-Pro, a conditional generative diffusion framework that enables efficient sampling of protein conformational trajectory space. TSS-Pro takes the initial frame as conditional input and generates protein conformational trajectories. It supports two sampling strategies: (1) consecutive sampling, where each trajectory segment is generated step by step by conditioning on the final frame of the previously generated segment, enabling temporally coherent propagation of structural transitions; (2) parallel sampling, where multiple trajectory branches are independently generated from initial conditions to enhance conformational diversity. We validate TSS-Pro on three representative systems of increasing complexity: alanine dipeptide, ubiquitin, and *Drosophila* cryptochrome (*d*CRY). TSS-Pro reproduces the free energy landscape of alanine dipeptide. In the case of ubiquitin, consecutive sampling with TSS-Pro overcomes local minima and uncovers distinct conformational states of the C-terminal region. For the large protein *d*CRY, TSS-Pro achieves high efficiency through parallel trajectory sampling, enabling conformational and dynamic exploration typically accessible only through extensive simulations. TSS-Pro paves the way for high-throughput exploration of protein trajectories and conformational landscapes for large and complex systems.

1. INTRODUCTION

Proteins, as fundamental molecular workhorses of life, underpin virtually all biological processes, supporting organismal growth, reproduction, and homeostasis.[1,2] The diverse biological functionalities of proteins depend on the precise folding of the amino acid sequences into three-dimensional (3D) structures. The relationship between protein structure and function is fundamental to life,[3−5] orchestrating processes ranging from enzyme catalysis[6] and signal transduction[7] to molecular transport[8] and disease development.[9] Experimental methods including X-ray crystallography,[10] nuclear magnetic resonance (NMR) spectroscopy,[11,12] and cryo-electron microscopy (cryo-EM)[13,14] have been widely used to determine protein structures.

While 3D structures provide a foundation for understanding proteins, sampling their dynamics is essential to capture the conformational flexibility that enables their function. Comprehensive characterization of protein dynamics provides critical insights into mechanisms of catalysis, regulation, and molecular recognition. Computational approaches, including molecular dynamics (MD)[15] simulations, coarse-grained MD simulations,[16,17] and enhanced sampling techniques, enable exploration of protein functional states and their underlying conformational landscapes. Machine Learning-assisted Force Fields (MLFF) have emerged as promising tools to enhance sampling efficiency.[18,19] The AI^{2}BMD model was developed using the ViSNet graph neural network,[20] and a fragmentation method was employed to enable *ab initio* accuracy in all-atom biomolecular simulations.[21] Despite significant advances in computational methods, dynamical sampling of biomolecular systems remains challenging because biologically relevant

Received: September 18, 2025
Revised: December 12, 2025
Accepted: December 22, 2025
Published: January 2, 2026

https://doi.org/10.1021/acs.jctc.5c01579

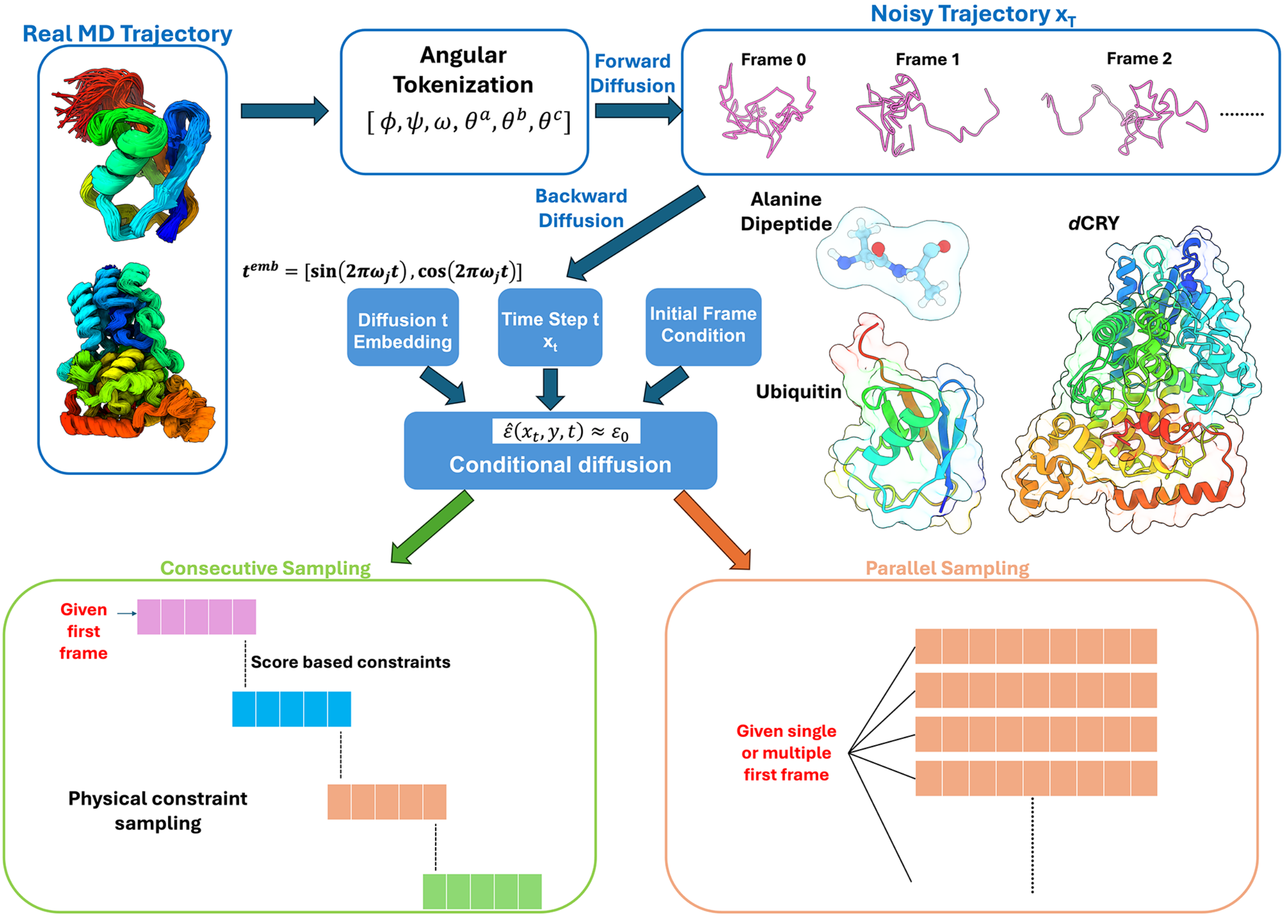


Figure 1. Architecture of TSS-Pro. Molecular dynamics (MD) trajectories were used as training data. During the backward diffusion process, the initial frame was incorporated as a conditional input to guide trajectory generation. The framework was evaluated on three systems: (1) alanine dipeptide; (2) ubiquitin; (3) *Drosophila* cryptochrome (*d*CRY). Two sampling strategies were implemented: consecutive sampling, which generates trajectories stepwise by refining the last frame of each segment and using it as the starting condition for the next segment, ensuring temporal continuity; and parallel sampling, which generates multiple trajectory segments independently from one or more reference frames, thereby increasing structural diversity and accelerating exploration of conformational space.

processes often occur on long time scales, ranging from microseconds to milliseconds, and even seconds.

The recent revolution in protein sequence-to-structure prediction, led by AlphaFold,[22−24] has significantly advanced protein conformational prediction and sampling. Since its release in 2020, AlphaFold has advanced from a convolutional neural network (CNN) architecture to a diffusion-based framework with substantial improvements in predictive accuracy.[24] Meta launched ESM Metagenomic Atlas, employing large language models to predict protein structures from primary sequence data.[25] Baker and coworkers developed RoseTTAFold, a generative diffusion model designed to address a broad range of protein design challenges, including *de novo* binder design, active-site scaffolding, and the engineering of functionally specialized protein structures.[26] These developments provide a strong foundation for protein conformational space sampling.

To bridge the gap between protein structural prediction and functional understanding, Frank Noé's team developed BioEmu,[27] which identifies key conformational changes such as the open-closed motions of Adenylate Kinase[28] and the conformational transitions of loops in Tetraspanin CD9.[29] Taking advantage of prevalent conditional diffusion model, Xu and Wang integrated protein state classifier, membrane constraint, and sequence conditioner as conditioning factors to guide the sampling of biologically relevant conformations of P-type ATPases.[30] Beyond discrete conformational sampling, other researchers have shifted their attention toward capturing the continuous protein trajectories and the transition pathways that connect distinct functional states. Liu et al. developed a deep learning framework following out-of-distribution detection methods for transition state identification based on MD simulations.[31] Ingólfsson et al. developed an automated multiscale framework incorporating machine learning−driven sampling and feedback, demonstrating its ability to efficiently capture lipid−protein interactions and conformational dynamics of the RAS−RAF signaling complex.[32]

Recently, a denoising diffusion probabilistic model (DDPM) has emerged as a powerful generative tool for enhancing the exploration of biomolecular conformational space. However, it still faces challenges, including underrepresentation of low-probability states and occasional production of physically implausible structures.[33] Extending this approach, Stirnemann and colleagues integrated solute-tempered Hamiltonian replica exchange with DDPMs, enabling markedly improved sampling across effective temperatures.[34] Despite such strengths in constructing conformational ensembles, efficient exploration of short-timescale protein dynamics remains an open challenge.

Given the strong performance of generative models in exploring protein dynamics and conformational space, we hypothesize that such models can be extended to sample protein trajectory space directly. This approach could enhance sampling efficiency with time-dependent information about conformational transitions often sought in MD simulations. Building on this idea, and drawing inspiration from the conditional diffusion model framework[35] and its implementation for protein backbone structure generation,[36] we introduce a conditional generative diffusion framework for sampling protein trajectory space, termed TSS-Pro. The framework supports two modes of implementation—consecutive and parallel—as illustrated in Figure 1. In consecutive sampling mode, trajectory segments are generated sequentially. Each new segment is derived from the final frame of the previous one, ensuring smooth and temporally consistent conformational evolution. In contrast, parallel sampling generates multiple trajectory segments from one or more reference frames, broadening the range of accessible conformations. To enhance sampling flexibility, TSS-Pro introduces a tunable pseudotemperature coefficient to modulate conformational variability.

To evaluate TSS-Pro, we selected three model systems of increasing size and complexity: alanine dipeptide, ubiquitin, and *Drosophila* cryptochrome (*d*CRY). For alanine dipeptide, TSS-Pro reconstructed a conformational landscape consistent with MD simulations. For ubiquitin, it overcame thermodynamic barriers and accessed additional conformational states beyond the MD ensemble. Finally, for *d*CRY, TSS-Pro scaled effectively to large proteins, generating physically plausible structures with a low atomic clash rate. Extending generative modeling to trajectory space, our work opens a new direction for molecular sampling in biomacromolecule systems.

2. METHODS

2.1. Molecular Dynamics Simulations

MD simulations were performed using the GPU-accelerated Amber22 package.[37,38] For *d*CRY, the AMBER ff14SB force field was applied to the protein, while the generalized AMBER force field (GAFF)[39] was used for the fully oxidized FAD cofactor, with RESP charges[40] assigned to its oxidation state. For ubiquitin and alanine dipeptide, the AMBER ff19SB[41] force field was applied. All simulations were carried out in explicit solvent using the TIP3P water model.[42] Covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm[43] to improve numerical stability. Energy minimization and equilibration protocols followed our previous work.[44] For ubiquitin and *d*CRY, the first 100 ns of the production trajectory were used for model training, and an additional 100 ns were used for analysis and conditional frame extraction, unless otherwise specified.

2.2. Data Preparation

The central idea of the TSS-Pro method is to combine sequential snapshots from MD trajectories to form a single training example. Therefore, each data point in both the training and generated data sets explicitly encodes temporal information about the system's dynamics.

We generated the training data by encoding each residue as six angular values: three torsion angles (Φ,ψ,ω) and three bond angles (θ^a,θ^b,θ^c), as described in a previous study.[36] The complete trajectory was divided into segments of 5, 10, 20, and 50 snapshots, respectively. In each partition, the first frame of every segment was designated as the reference structure $y \in \mathbb{R}^{L\times 6}$, where L denotes the number of residues, and the six angles are stacked along the second dimension in the order $\Phi,\psi,\omega,\theta^a,\theta^b,\theta^c$. The angular values of all subsequent frames were computed as differences from the reference frame. The remaining frames of each segment were represented as a two-dimensional array $x \in \mathbb{R}^{L\times 6k}$, where k is the number of predicted snapshots in the segment (e.g., k is 4 for segments with 5 snapshots), and L is the number of residues. As is depicted in Figure S12, frames were concatenated along the second dimension (e.g., [frame 1|frame 2| ... |frame k]), with each frame containing six angles. The reference frame of each segment x, denoted y, served as the conditioning input during training. This formulation yields angular trajectories computed relative to the initial frame, providing a compact and transformation-invariant representation well-suited for trajectory learning.

2.3. Augmented Dickey-Fuller (ADF) Test for Protein Trajectory Data

The ADF test[45,46] was performed on the protein trajectory data to test for stationarity. The general form of the ADF regression is

$$\Delta x_t = \mu + \gamma t + \alpha x_{t-1} + \sum_{j=1}^{k-1} \beta_j \Delta x_{t-j} + u_t \tag{1}$$

where $\Delta x_t = x_t - x_{t-1}$ denotes the first difference of the time-series data (the five temporal features mentioned in the time-series analysis for the protein trajectory section), μ is a constant, and u_t is white noise. The lagged differences Δx_{t-j} account for higher-order serial correlation, while the coefficients β_j capture the short-run dynamics of the differenced series.

The ADF function is computed as

$$ADF = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \tag{2}$$

where $\hat{\alpha}$ is the Ordinary Least Squares (OLS) estimate of α. This statistic tests the null hypothesis H_0: $\alpha = 0$ (presence of a unit root, nonstationarity) against the alternative H_1: $\alpha < 0$ (stationarity). When $\alpha > 0$, the process is explosive and nonstationary, with variance growing without bound. The associated p-values represent the probability of observing the test results under the null hypothesis; $p < 0.05$ suggests rejection of H_0, supporting the stationarity of the data.

2.4. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

The ACF and PACF functions are used to characterize temporal dependencies in protein trajectories. The ACF measures the correlation between observations and their lagged counterparts, quantifying the persistence of patterns over time. Given a time series with N observations, let $\bar{x}$ denote the sample mean. The autocorrelation at lag k is defined as

$$ACF(k) = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{N} (x_t - \bar{x})^2} \tag{3}$$

The PACF quantifies the direct relationship between x_t and x_{t-k} after removing the linear effects of all intermediate lags from 1 to $k - 1$. Formally, the PACF at lag k is given by the last coefficient Φ_{kk} in the autoregressive model of order k:

$$x_t = \phi_{k1}x_{t-1} + \phi_{k2}x_{t-2} + ... + \phi_{kk}x_{t-k} + \varepsilon_t \tag{4}$$

where ε_t represents the white noise error term. The coefficient Φ_{kk} corresponds to the partial autocorrelation at lag k. In this study, PACF values are estimated using the Yule-Walker equations with unbiased autocovariance estimates,[47] as implemented in the statsmodels package.[48]

2.5. Conditional Generative Diffusion Model in TSS-Pro

To generate protein trajectories, we adopted a conditional Denoising Diffusion Probabilistic Model (cDDPM), extending the framework of the previously reported unconditional FoldingDiff model.[36] Each segment (as described in Section 2.2) serves as the ground-truth sample in the diffusion model. During training, the model learns a Markovian forward noising process in which Gaussian noise is gradually added to a ground-truth sample, denoted x_0, over a total of T steps. Simultaneously, it learns a corresponding reverse denoising process, parametrized by a trainable neural network, that progressively reconstructs x_0 from the noisy inputs. The overall objective is to model $p_\theta(x_{0:T}|y)$, where $x_{0:T} = \{x_0, x_1, x_2, ..., x_T\}$ denotes the sequence of data points in the diffusion process, with progressively increasing noise at each step. The structure of each data point in $x_{0:T}$ follows the same format as described in Section 2.2, *Data preparation:* the reference frame $y \in R^{N\times 6}$, corresponding to x_0, serves as the conditioning input to train the denoising network.

During the forward diffusion process, at each step $t \in \{0,1,2,...,T\}$, the protein trajectory as segment data x_{t-1} is corrupted by Gaussian noise to produce x_t, defined as

$$q(x_t|x_{t-1}) = \mathcal{N}_{wrapped}(x_t; x_{t-1}\sqrt{1-\beta_t}, \beta_t \boldsymbol{I}) \tag{5}$$

where β_t refers to the variance schedule at step t, and $\mathcal{N}_{wrapped}$ indicates that angular values are wrapped into the principal interval range $(-\pi,\pi]$. Using the closed-form expression for the marginal distribution $q(x_t|x_0)$, the forward process transformation is

$$x_t = wrap(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon),\ \varepsilon \sim \mathcal{N}(0, \boldsymbol{I}) \tag{6}$$

where $\bar{\alpha}_t$ is derived from the cosine schedule, similar to β_t.[36,49] The wrapping function ensures angles remain within the principal interval:

$$wrap(X) = [(X + \pi)\mathrm{mod}2\pi] - \pi \tag{7}$$

The diffusion model is trained to predict the noise term ε, as defined in eq 6, rather than directly reconstructing x_t. The neural network $\hat{\varepsilon}(x_t, y, t)$ is conditioned on three inputs: the noised protein trajectory x_t, the reference frame y, and the diffusion step t. The loss function for training is defined following the previous study:[36]

$$l_{SmoothL1,wrapped} = \begin{cases} \dfrac{d_w^2}{2\eta} & |d_w| < \eta \\ |d_w| - 0.5\eta & otherwise \end{cases},\ d_w = wrap(\varepsilon - \hat{\varepsilon}(x_t, y, t)) \tag{8}$$

where $\eta = 0.1\pi$ denotes the transition threshold between the L2 and L1 loss regimes. In the L1 regime, large errors are penalized linearly, reducing sensitivity to outliers, whereas in the L2 regime, small errors are penalized quadratically, encouraging precision.

In the reverse diffusion process, the model iteratively denoises the trajectory x_t to reconstruct x_{t-1}, conditioned on the reference frame y. The reverse transition distribution is defined as

$$q_\theta(x_{t-1}|y, x_t) = \mathcal{N}_{wrapped}(x_{t-1}; \mu_t(x_t, y, t), \sigma_t^2\boldsymbol{I}) \tag{9}$$

The posterior mean μ_t and variance σ_t^2 are determined by the noisy trajectory x_t, the conditioning reference frame y, and the diffusion step t:

$$\mu_t(x_t, t, y) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\varepsilon}(x_t, y, t)\right) \tag{10}$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{11}$$

At each reverse step, x_{t-1} is sampled as

$$x_{t-1} = wrap(\mu_t(x_t, y, t) + \zeta\sigma_t z) \quad z \sim \mathcal{N}(0, \boldsymbol{I}) \tag{12}$$

where ζ is a stochastic coefficient that controls the degree of randomness during sampling. A setting of $\zeta = 1$ yields fully stochastic sampling, whereas $\zeta = 0$ produces a deterministic trajectory, similar to Denoising Diffusion Implicit Models (DDIMs).[50] Here, $z \sim \mathcal{N}(0, \boldsymbol{I})$ represents independent standard Gaussian noise, which injects stochasticity into each step when $\zeta > 0$. The process is initialized by drawing $\mathcal{N}(0, \boldsymbol{I})$, and eq 12 is then applied iteratively in reverse order, from $t = T$ down to $t = 1$, to progressively denoise the trajectory and reconstruct x_0. During the denoising process, x_T is generated via Gaussian noise sampling. At each step $t \in [0,T]$, x_{t-1} is computed by the Denoising Diffusion Probabilistic Model (DDPM), using the noise term $\hat{\varepsilon}$ predicted by the BERT network, as defined in eq 12. This iterative process continues until an estimate of the clean sample x_0 is obtained. Note that the index t refers to the diffusion time step, not the physical time step index in the MD trajectory. A denoising diffusion process with $T = 1000$ is illustrated in Figure S3 of the Supporting Information. It starts with angular values sampled from Gaussian noise at diffusion time $t = 1000$ and gradually regains protein folding and conformational plausibility as t decreases to 0.

2.6. Neural Network Architecture

A BERT-based Transformer encoder[51] was employed as the learnable denoising network. The noisy trajectory x_t, the conditioning reference frame y, and the diffusion time step t (embedded using random Fourier features) are projected into a shared 384-dimensional space and then summed to form the input representation:

$$h = x_t^{emb} + y^{emb} + t^{emb} \in \mathbb{R}^d \tag{13}$$

The BERT encoder was implemented with 12 Transformer layers, each containing 12 attention heads, following previous studies (Table S4).[36,52] This architecture integrates temporal, structural, and conditional information and provides output $\hat{\varepsilon}(x_t, y, t)$ with the same dimensionality as x_t.

2.7. PyRosetta Optimization

As a postprocessing step during sequential sampling, the last frame of each trajectory segment was refined with PyRosetta[53] using harmonic coordinate constraints applied to backbone heavy atoms. This refinement improved local geometry, reduced steric clashes, and preserved the overall backbone conformation. The all-atom Rosetta Energy Function 2015 (REF15)[54] was applied in conjunction with a constraint-weighted relaxation protocol, where harmonic coordinate constraints were gradually relaxed over 1,000 iterations.

2.8. Latent Space Distribution and Clustering Analysis

To visualize the protein conformational landscape, we employed two approaches for comparison: a feedforward autoencoder[55] and time-lagged independent component analysis (t-ICA).[56] Following our previous study,[44] the autoencoder architecture was implemented with an encoder and a decoder, using Rectified Linear Unit (ReLU) activations in all hidden layers and a sigmoid output layer to constrain reconstructions within the [0,1] range. The network was trained by minimizing the mean-squared reconstruction error, augmented with an additional L2 penalty on the latent variables to favor a compact and well-structured latent space. t-ICA was implemented using the PyEMMA 2 package[57] following standard protocols, and the resulting trajectories were projected onto the first two time-lagged independent components to yield a two-dimensional kinetic embedding.[58]

Protein trajectory conformations were clustered using the MDSCAN package,[52] which is designed to analyze long MD trajectories based on the root-mean-square deviation (RMSD). MDSCAN enhances the HDBSCAN algorithm by introducing a memory-efficient RMSD-based implementation that uses vantage-point tree encoding to speed up nearest-neighbor searches and a dual-heap approach to build a quasi-minimum spanning tree. HDBSCAN[59] is a hierarchical density-based clustering algorithm that identifies clusters of varying densities by building a hierarchy of connected regions in data. Compared with HDBSCAN, MDSCAN dramatically reducing time and memory complexity for clustering long molecular dynamics trajectories.

2.9. Conversion between Backbone Angles and 3D Coordinates

To represent a protein backbone in a compact yet informative manner, each residue is encoded using six internal angles: three torsion angles $(\Phi,\psi,\omega)$and three bond angles $(\theta^a,\theta^b,\theta^c)$, as illustrated in Figure S7 of the Supporting Information. For a protein of L residues, this yields an angle matrix of dimensions equal to $L \times 6$. However, the first residue lacks the Φ torsion angle, and the terminal residue lacks ψ, ω, θ^b, and θ^c. The values of these angles are initialized to zero during training and excluded during the reconstruction of the protein structure from the output. The reconstruction of Cartesian coordinates from these internal coordinates follows the Natural Extension Reference Frame (NeRF) algorithm.[60] Given three backbone atoms in residue i: N^i, C_α^i, $C^i \in R^3$, the position of the next backbone atom N^{i+1} can be determined using the NeRF algorithm. Iteratively applying this transformation reconstructs the backbone coordinates following the repeating atomic order $N-C_\alpha-C$. This formulation ensures that the protein structure can be faithfully regenerated from the internal angle representation while maintaining geometric continuity. To evaluate reconstruction accuracy, RMSD values were calculated for alanine dipeptide, ubiquitin, and dCRY by comparing structures before and after reconstruction. The results are reported in Table S2 of the Supporting Information. The consistently low RMSD values indicate that the reconstructed structures closely match the original conformations, thereby confirming the reliability of the reconstruction process.

2.10. Potential of Mean Force (PMF)

The PMF is the free energy profile of a system as a function of a chosen reaction coordinate or collective variable,

$$W(z) = -k_B T \ln P(z) \tag{14}$$

where k_B is the Boltzmann constant, T is the ambient temperature 300 K, and $P(z)$ is the probability distribution of the conformation at bin z. In this study, the configuration space is discretized into 100 uniform bins along each dimension.

2.11. Kullback–Leibler (K–L) Divergence

To quantify the deviation between two probability distributions, we computed the K–L Divergence following established practices in prior work. Given two discrete distributions P and Q, the K–L Divergence is defined as:

$$D_{KL}(P\|Q) = \sum_i^N P_i \log\left(\frac{P_i}{Q_i}\right) \tag{15}$$

Where N denotes the number of histogram bins used to represent the distributions. A lower D_{KL}value indicates greater similarity between P and Q, while higher values reflect increased divergence.

2.12. Signal-to-Noise Ratio (SNR) Calculation

To quantitatively evaluate the overlap between the $\phi-\psi$ angle distributions sampled by TSS-Pro and MD, we computed the Signal-to-Noise Ratio (SNR)[61,62] based on the popularity density over $\phi-\psi$ space, as given by the following equation:

$$\mathrm{SNR}_{dB} = 10\log_{10}\left(\frac{\langle P_{\mathrm{MD}}^2\rangle}{\langle (P_{TSS-Pro} - P_{MD})^2\rangle}\right) \tag{16}$$

where $\langle\cdot\rangle$ denotes the mean over all histogram bins in the $\phi-\psi$ dimension space. $P_{TSS-Pro}$ and P_{MD} denote the probability densities obtained from sampling with TSS-Pro and from MD simulation, respectively. The MD distribution is taken as the signal, and the deviation of the TSS-Pro distribution from it is considered the noise.

3. RESULTS AND DISCUSSION

3.1. Overall Architecture

Training of the conditional diffusion framework was performed using molecular dynamics trajectories from three model systems (Figure 1). During the forward diffusion step, Gaussian noise was added to the trajectory based on the eq 6, transforming it into the noisy trajectory x_T. The model was trained to predict the noise by $\hat{\varepsilon}(x_t, y, t)$.

During inference, the initial frame y is provided as a condition, together with the diffusion time embedding t^{emb} = $[\sin(2\pi\omega_j t),\cos(2\pi\omega_j t)]$ and the protein trajectory angular values x_t at diffusion time t. A transformer-based network then predicts the noise and iteratively denoises the trajectory, reconstructing physically meaningful conformational dynamics.

We implemented two sampling strategies. Consecutive sampling—illustrated in the green panel (bottom left) of Figure 1—is initialized from a single conformation and generates the protein trajectory one segment at a time. The final frame of trajectory segment i is refined using PyRosetta, with harmonic coordinate constraints applied to backbone

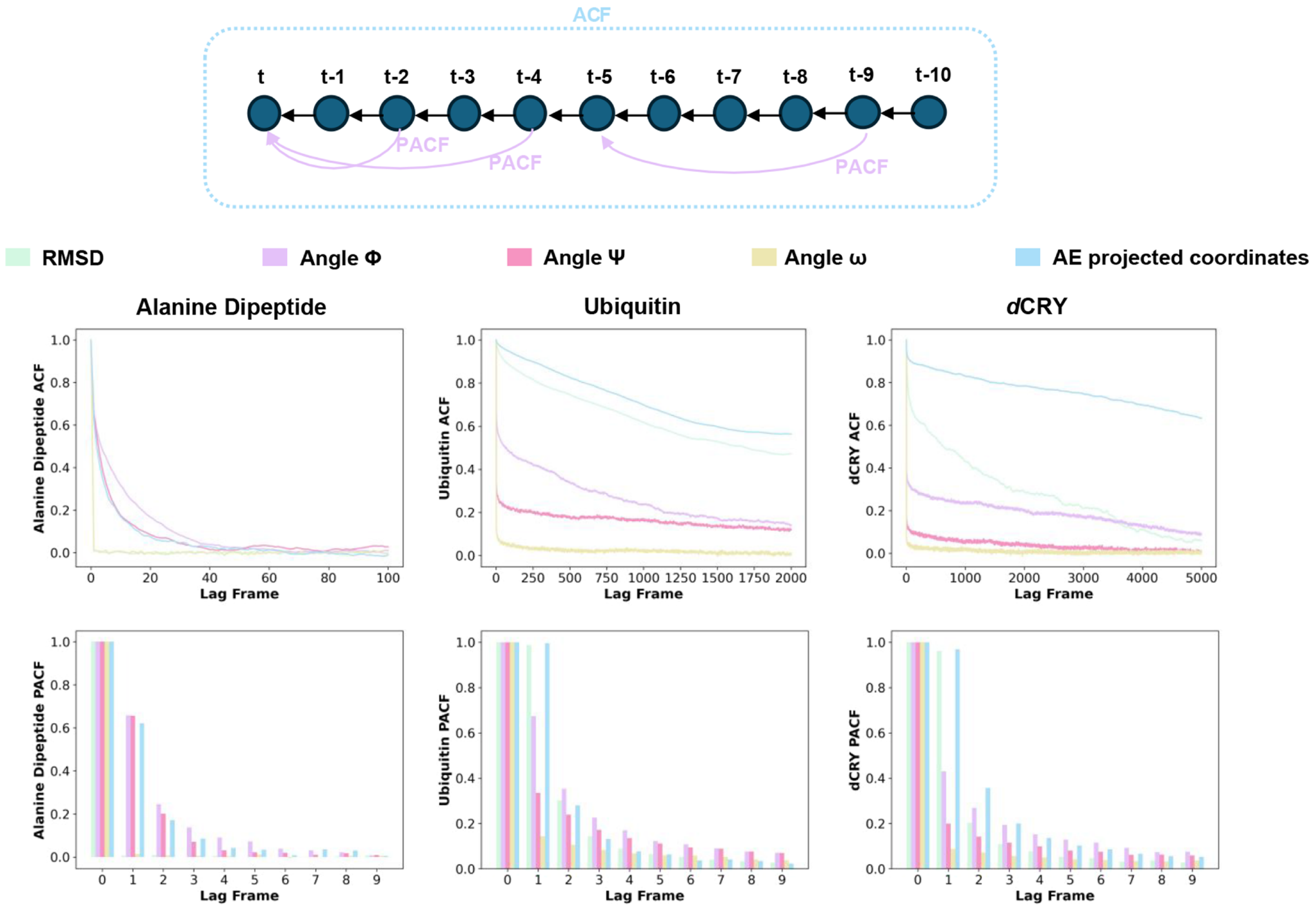


Figure 2. Autocorrelation function (ACF) and partial autocorrelation function (PACF) for alanine dipeptide, ubiquitin, and *d*CRY trajectory data. Temporal features analyzed include: (1) Root-mean-square deviation (RMSD, mint green), quantifying structural deviation relative to the initial frame; (2) Mean backbone dihedral angle Φ (torsion around peptide bond N−Cα, lavender); (3) Mean backbone dihedral angle ψ (torsion around peptide bond Cα−C, pink); (4) Mean dihedral angle ω (torsion around the peptide bond C−N, Beige); and (5) One-dimensional protein coordinate representations derived from the autoencoder (AE) model (sky blue). The ACF measures both direct and indirect correlations between a variable and its past values, providing an overall assessment of temporal dependence across lag times. In contrast, the PACF isolates the direct contribution of each lag after accounting for the effects of shorter lags.

atoms during optimization. This produces a physically plausible structure that serves as the starting condition for the subsequent segment $i + 1$. This approach ensures temporal continuity across the generated trajectories. In contrast, parallel sampling—illustrated in the orange panel (bottom right) of Figure 1—generates trajectory segments independently from one or more reference frames. This design supports rapid diversification of sampled conformations through concurrent trajectory expansion, resembling a breath-first search (BFS) strategy.[63]

3.2. Time-Series Analysis for the Protein Trajectory

To investigate temporal dependencies in protein trajectories, we conducted time series analysis on the generated data. Because both autocorrelation function (ACF) and partial autocorrelation function (PACF) assume stationarity, we first tested the trajectory data to confirm this condition. Five temporal features were then extracted as functions of time: (1) RMSD calculated relative to the first frame; (2) Mean backbone dihedral angle Φ (torsion around peptide bond N−Cα); (3) Mean backbone dihedral angle ψ (torsion around peptide bond Cα−C); (4) Mean dihedral angle ω (torsion around the peptide bond C−N); (5) One-dimensional protein coordinate representations derived from the autoencoder (AE) model. Dihedral angles were averaged across all residues. The backbone dihedral angles are depicted in angular tokenization process in Figure S7.

To clarify the stationarity of the five time-dependent features, we applied the Augmented Dickey-Fuller (ADF) test. The ADF statistics and corresponding p-values are reported in Table S1 of the Supporting Information. Across all three systems, alanine dipeptide, ubiquitin, and *d*CRY, the five features consistently exhibited negative ADF statistics and significantly low p-values (<0.05), demonstrating their stationarity. Notably, the ADF statistics for alanine dipeptide are substantially lower (i.e., more negative) than those of ubiquitin and *d*CRY, indicating a higher degree of stationarity in this smaller dipeptide system. This stationarity supports the validity of subsequent time series analyses using ACF and PACF.

After verifying stationarity, we calculated the ACF and PACF coefficients. The ACF captures both direct and indirect correlations between observations at different time lags; for example, the effect of x_{t-2} on x_t may be mediated through x_{t-1}. In contrast, the PACF measures the direct linear relationship between x_t and x_{t-k} after accounting for the influence of all intermediate lags by conditioning on the values of

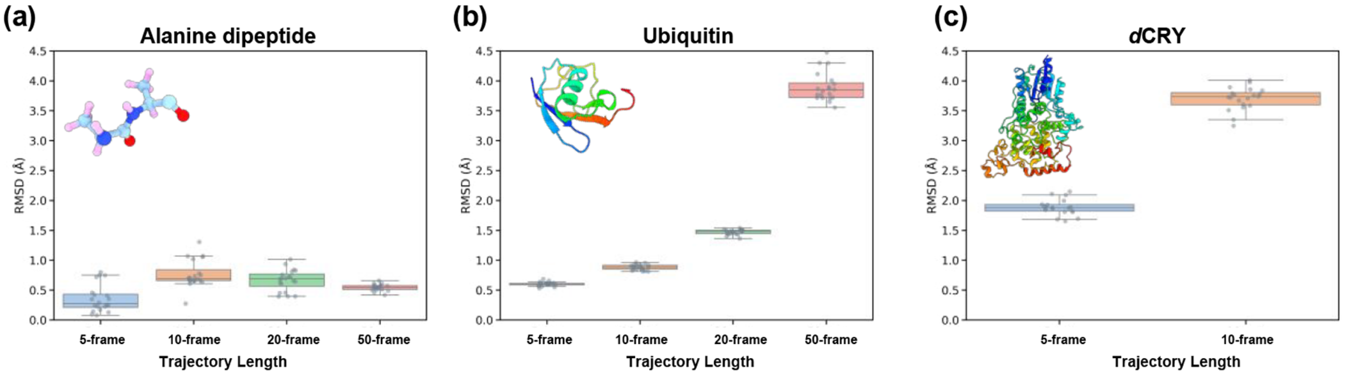


Figure 3. Average RMSD values for generated trajectories of varying lengths for (a) alanine dipeptide, (b) ubiquitin, and (c) *d*CRY, computed from 20 generated segments per system.

$x_{t-1}, x_{t-2}, ..., x_{t-k+1}$. As formalized in eq 4, the coefficient Φ_{kk} denotes the partial autocorrelation at lag k.

Both ACF and PACF coefficients for all five time-dependent protein trajectory features decrease with increasing lag time (Figure 2). Among these features, the AE-projected coordinates, RMSD, and dihedral angle Φ exhibit a slower decay rate compared to the other two features. For alanine dipeptide, the ACF and PACF coefficients decline much more rapidly with lag time than those of the larger proteins. Across all the three systems, PACF coefficients fall below 0.2 within five steps and below 0.1 within ten steps, indicating that protein dynamics are predominantly influenced by the preceding 5−10 frames. This faster decay in smaller systems reflects their shorter memory and faster relaxation dynamics, whereas larger proteins exhibit longer temporal correlations due to their greater structural complexity.

3.3. Trajectory Generation by TSS-Pro

Since the protein conformations are predominantly influenced by the preceding 5−10 frames, we focused on the 5-frame and 10-frame segment models for all three model systems, which correspond to sequences of a reference frame followed by 4 or 9 subsequent frames, respectively. The performance of the conditional diffusion model was assessed based on the average RMSD of frames in each segment relative to the reference frame (Figure 3). For alanine dipeptide, the average RMSD relative to the first frame remained largely unchanged with increasing trajectory length, consistent with the PACF results (Figure 2), which show that correlations vanish after approximately five steps. In contrast, ubiquitin and *d*CRY exhibited increasing RMSD with longer trajectories, reaching approximately 3.5−4 Å for *d*CRY in the 50-frame trajectory segments. These results indicate that while the model captures short-term dynamics reliably (4−5 frames ahead), its accuracy declines over longer horizons due to error accumulation and the broader conformational space of larger proteins.

As shown in Figure 4a, RMSD trends for five representative model-generated trajectories from each system indicate that ubiquitin exhibits a clear monotonic increase in RMSD with frame number in both the 5-frame and 10-frame segment models, suggesting the presence of time-correlated deviations. In contrast, *d*CRY displays weaker and less consistent trends. Representative structural snapshots in Figure 4b reveal modest conformational differences between generated frames, consistent with the limited structural changes expected for protein dynamics at the picosecond time scale.

To further assess the role of the stochastic coefficient ζ in eq 12, we analyzed the frame-dependent RMSD profiles of the ubiquitin system under different ζ values (Figures S1 and S2). At very small values (e.g., $\zeta = 0.005$), the sampled trajectories showed minimal variance, yielding nearly identical RMSD plots across batches. As ζ increases, the trajectories show increasing diversity, with larger variability in RMSD values. This behavior reflects the influence of the additive Gaussian noise term $\zeta\sigma_t z$ in eq 12, which injects increasing randomness into the sampling process at higher ζ. At $\zeta = 0.1$, the variability among trajectories closely resembled that observed in real MD simulations, particularly in the distribution of RMSD profiles (Figure 4a). These results demonstrate that our conditional diffusion framework provides tunable control over trajectory diversity, enabling the generation of either highly consistent or highly varied conformational pathways by tuning ζ.

3.4. Sampling the Free Energy Landscape of Alanine Dipeptide

To evaluate the performance of the conditional diffusion model in exploring conformational space, we compared its sampled conformations with those from the corresponding MD simulations across the three systems. For alanine dipeptide, we computed the $\Phi - \psi$ PMF based on the MD simulation and generated trajectories using a consecutive sampling scheme with segment lengths of 5, 10, and 20 frames. To match the length of the 100 ns MD simulations used in this analysis, a total of 100 ns of trajectories was generated for each scheme. The 5-frame segment model produced the more rigid landscape (Figure 5b) than MD simulation (Figure 5a). In contrast, the 10-frame and 20-frame models exhibited greater dispersion in conformational space. Overall, the model-generated and simulated 100 ns trajectories showed consistent trends, with high-density regions concentrated within $\Phi \in [-175°, -25°]$ and $\psi \in [-100°, 150°]$. Although the PMF based on trajectories generated by the conditional diffusion model does not fully overlap with the free energy surface (FES) derived from density functional theory (DFT),[64] it still captures much of the energetically favorable region in $\Phi \in [-180°, -60°]$ and $\psi \in [-100°, 150°]$. Furthermore, these PMFs align well with a previous study.[65]

To quantify the overlap between the MD simulated ensemble and consecutively sampled ensemble, we calculated the K−L divergence, as described in Section 2.11. The K−L divergence is much lower in the 5-, 10-, and 20-frame segment schemes compared to the 50-frame segment schemes (Table S3 of the Supporting Information). This observation

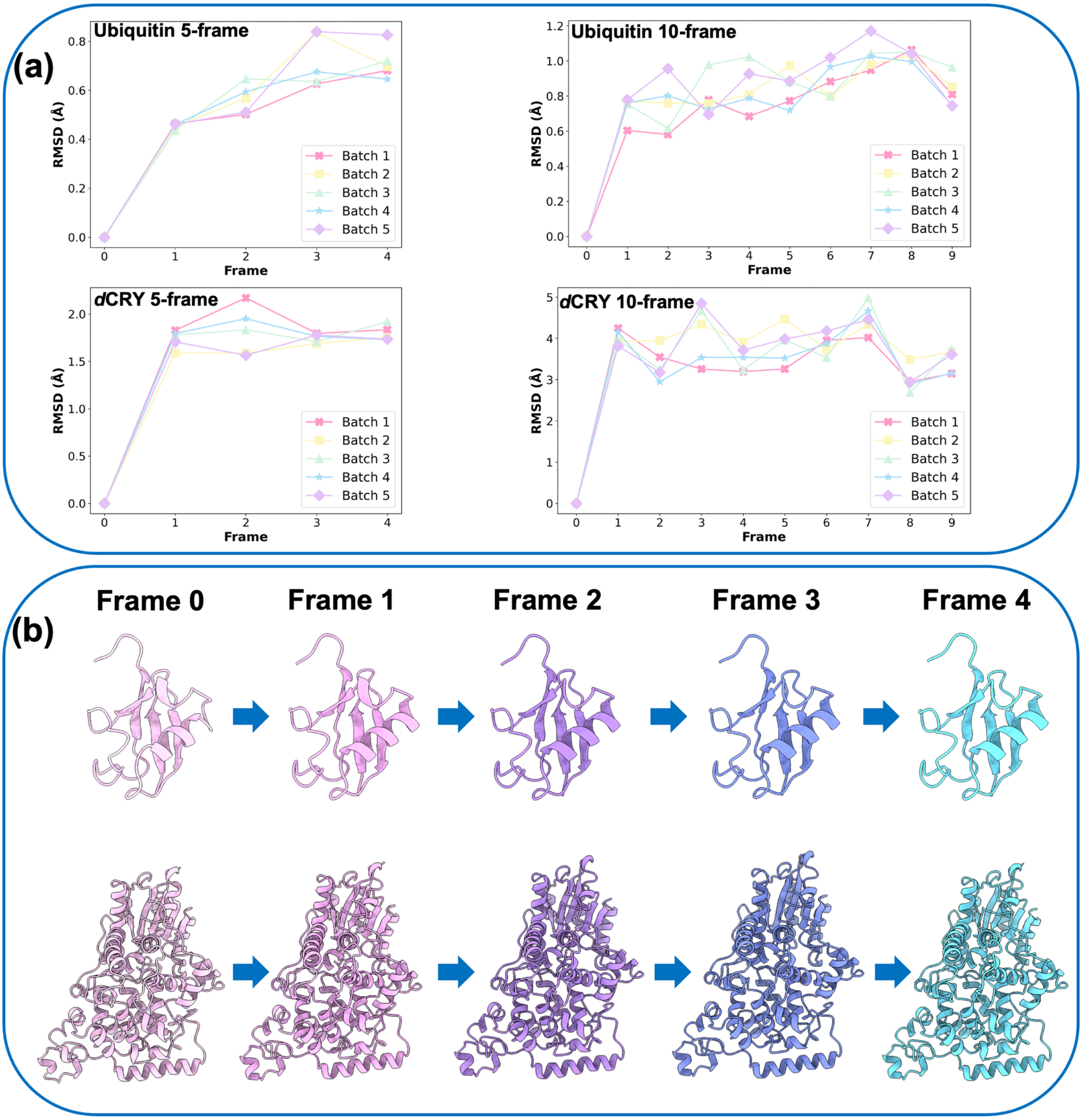


Figure 4. (a) RMSD as a function of frame number for the generated trajectories with 5 different batches. Colored lines represent individual trials, highlighting run-to-run variability and the diversity of conformations generated by TSS-Pro. (b) Representative structure snapshots from the generated trajectories, illustrating conformational changes over time (top: ubiquitin; bottom: *d*CRY).

demonstrates the capability of the new method developed in this study to reliably reproduce the free energy landscape of model systems obtained through conventional simulations.

Additionally, the Signal-to-Noise Ratio was used to quantify the degree of overlap between conformational distributions sampled by TSS-Pro and those from MD simulations. As shown in Supplementary Figure S14, the SNR remains consistently above zero, indicating that TSS-Pro sampling captures key characteristics of the MD-derived distribution. It increases moderately as the TSS-Pro segment length increases from 5 to 20 frames, reaching a maximum of 3.98 dB. This trend suggests improved agreement with MD simulations as the segment length increases. However, under the 50-frame segmentation scheme, the SNR drops sharply to 0.97 dB, indicating that the model may have limited capacity to accurately sample long trajectories. While further evaluation is needed to establish clear implementation guidelines, this falls outside the scope of the present proof-of-concept study.

3.5. Exploration of Ubiquitin Conformational Space Using TSS-Pro

3.5.1. Consecutive Sampling Model. We first applied consecutive sampling to ubiquitin using TSS-Pro to generate a

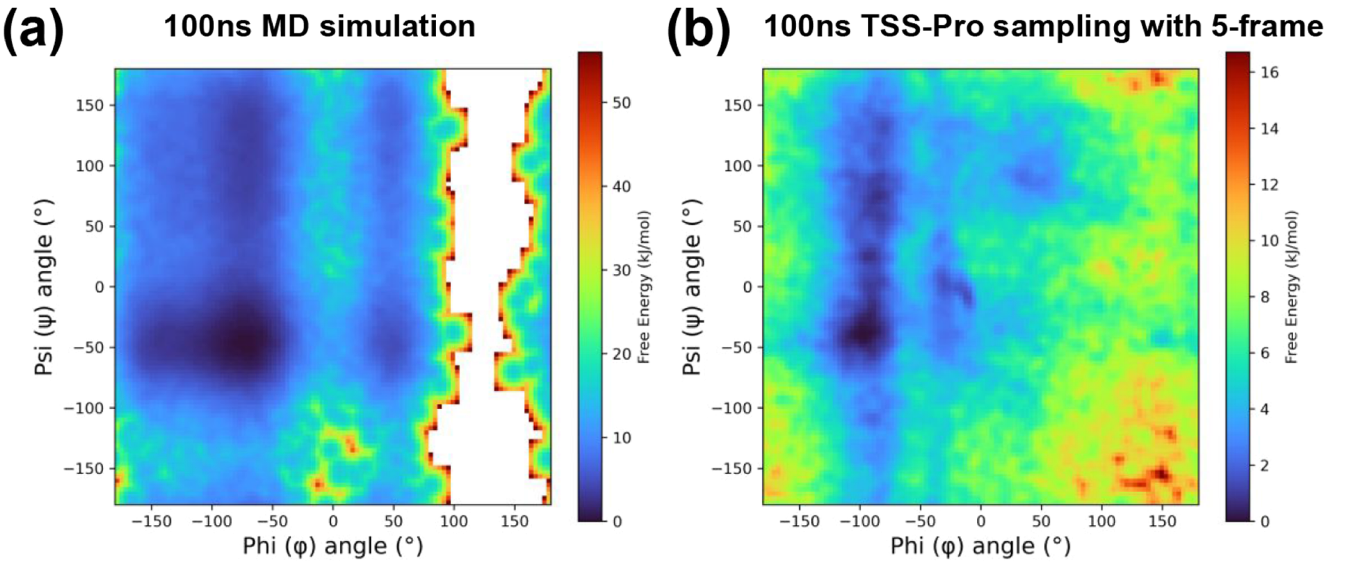


Figure 5. Potential of Mean Force (PMF) profiles of the ϕ-ψ dihedral angle distribution for alanine dipeptide: (a) 100 ns MD simulation, (b) TSS-Pro sampling through 5-frame model, (c) TSS-Pro sampling through 10-frame model, and (d) TSS-Pro sampling through 20-frame model.

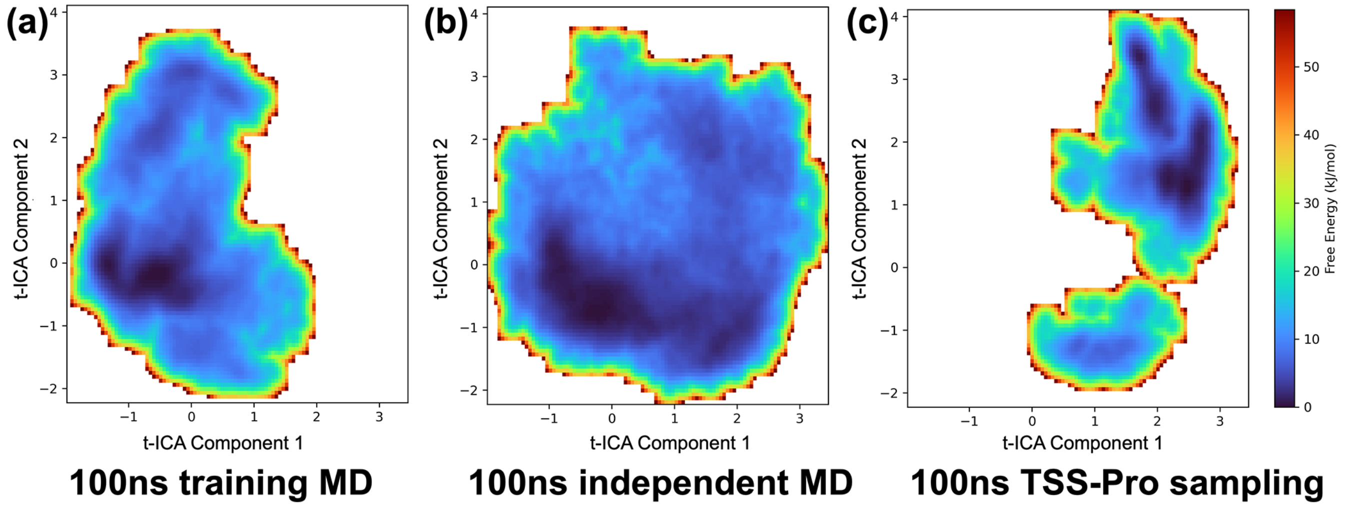


Figure 6. Comparison of the conformational space distributions from MD simulations and TSS-Pro generated trajectories. Potential of Mean Force of (a) 100 ns MD used for training, (b) 100 ns independent MD, and (c) 100 ns ubiquitin TSS-Pro consecutive sampling.

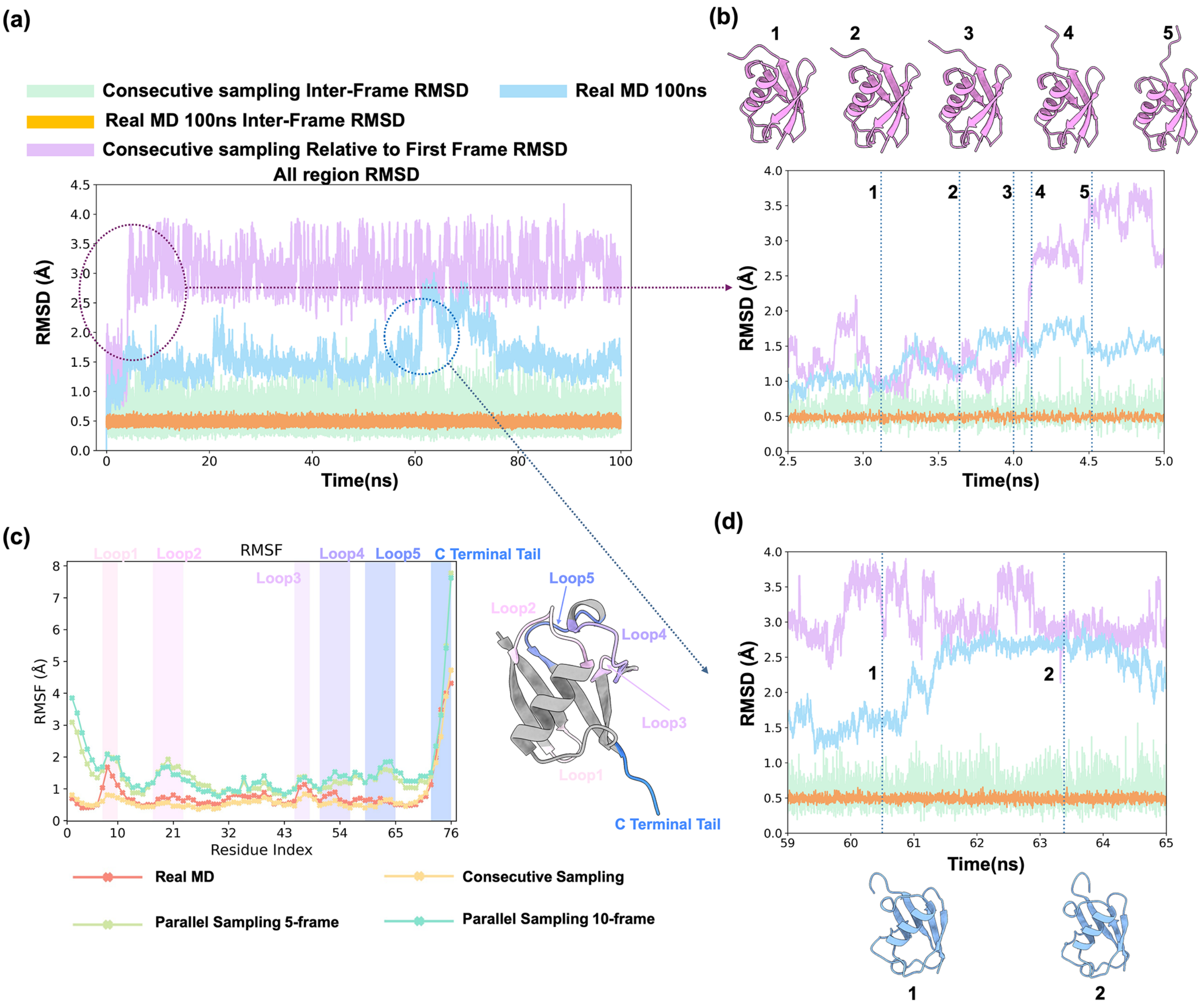


Figure 7. (a) RMSD evolution for predicted trajectories and MD simulations over 100 ns, demonstrating consistency with MD dynamics. Close-up view for the jump of RMSD over (b) TSS-Pro-generated trajectory and (d) 100 ns training MD trajectory, and (c) Root-mean-square fluctuation (RMSF) profiles comparing TSS-Pro−generated trajectories with MD simulations, highlighting flexible loop regions and the C-terminal tail (CTT). The RGG (74−76) motif shows the highest fluctuations, consistent with its functional role.

100 ns trajectory. In addition to the 100 ns MD simulation used as training data to develop the conditional generative diffusion model, an independent 100 ns MD simulation was generated for comparison. Following a previous study,[66] both MD simulations were used to build a t-ICA latent space from which the potential of mean force was constructed for three sets of ubiquitin trajectories: the training simulation, the independent simulation, and 100 ns of consecutive samples generated using TSS-Pro (Figure 6). The independent simulation (Figure 6b) explores a broader region of the latent space than the training simulation (Figure 6a). Interestingly, the TSS-Pro consecutive sampling (Figure 6c) primarily occupies the region sampled by the independent simulation, rather than by the training simulation.

Further structural analysis reveals that the differences among these sampling results mainly originate from the C-terminal tail (CTT), which is known to be flexible and critical for ubiquitin function.[67] We present the RMSD of the ubiquitin core structure, comprising residues 1−73 and excluding the CTT, for all three samplings in Figure S9 of the Supporting Information. As the RMSD of the ubiquitin core structure remains under 1.75 Å throughout all the sampling data, this indicates that the TSS-Pro-generated sampling does not simply replicate either simulation but instead yields complementary sampling data, which is an expected outcome of independent simulation methods. This is further supported by the observation that the K−L divergence between the TSS-Pro-generated trajectory and the independent simulation is smaller than its divergence from the training simulation (Table S7). Additionally, we performed Markov State Model (MSM) analysis and computed the implied time scales for both TSS-Pro and MD trajectories. The results show comparable relaxation behavior (Figure S10) and metastable states (Figure S11), confirming that the TSS-Pro−sampled trajectories preserve the temporal correlations and kinetic properties observed in MD simulations.

To evaluate the stability of consecutive sampling, we computed both the RMSD between each frame and its

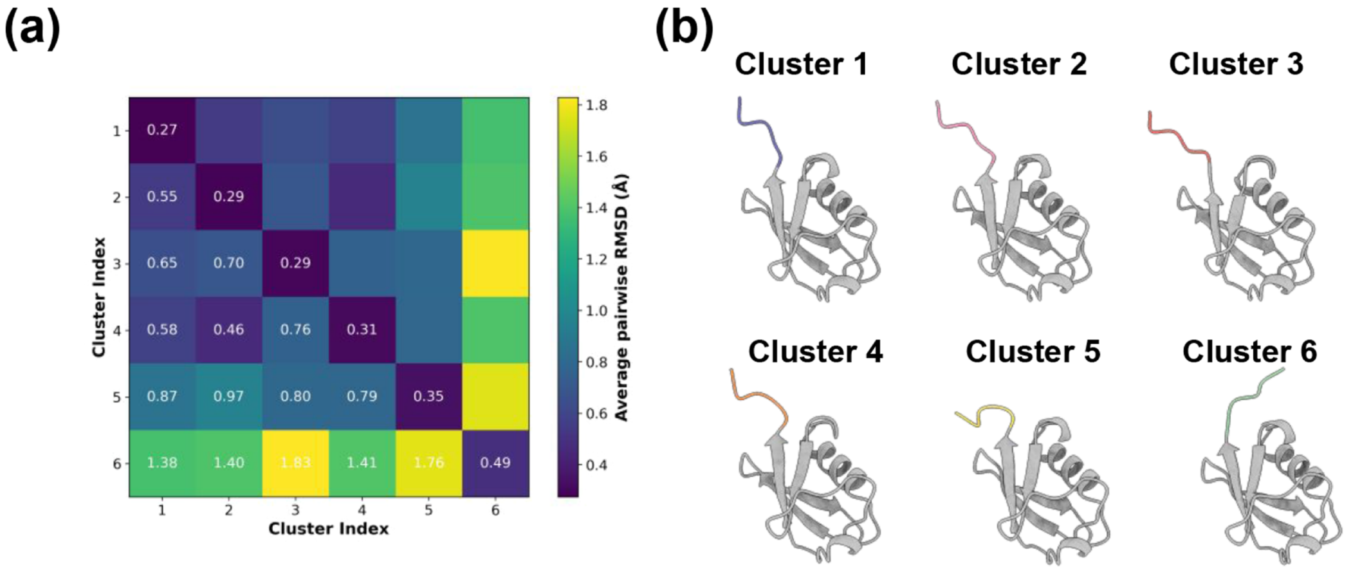


Figure 8. MDSCAN-Based Clustering of Ubiquitin Conformations from TSS-Pro Consecutive Sampling. (a) Average pairwise RMSD values between clusters; (b) Representative centroid structure of each clusters.

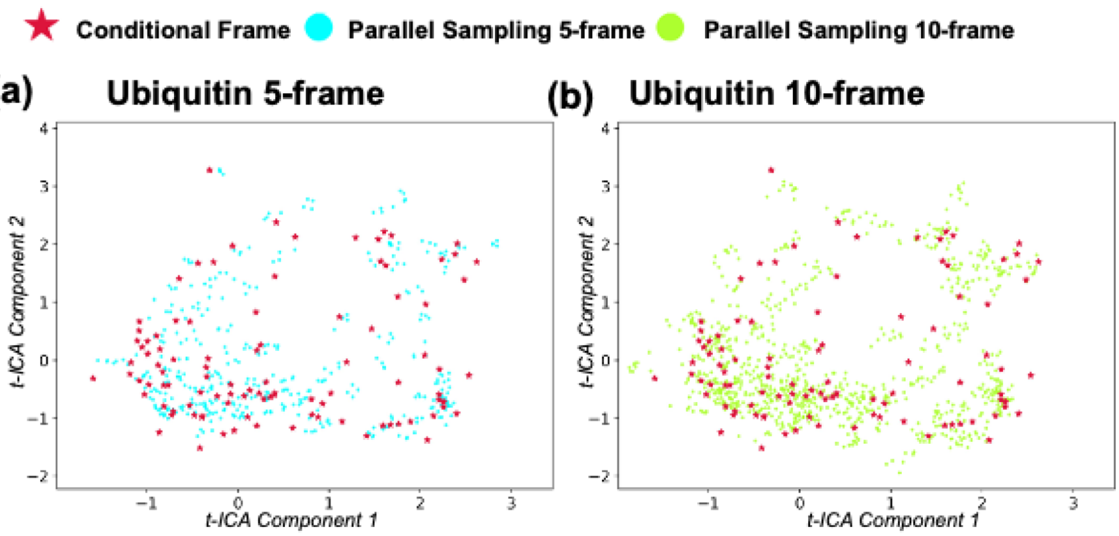


Figure 9. Parallel sampling distribution based on the conditional frames over (a) Ubiquitin 5-frame and (b) Ubiquitin 10-frame.

successor (adjacent-frame RMSD) and the RMSD with respect to the first frame (Figure 7a). The adjacent-frame RMSD values remain below 1 Å, indicating gradual and stable structural evolution throughout the trajectory. In contrast, the RMSD with respect to the first frame shows a significant increase around 4 ns. Structural analysis reveals that this is associated with a conformational transition, involving a change in the orientation of the ubiquitin CTT (residues 74−76) (Figure 7b). During this switch, the ubiquitin core structure remains stable. This transition also corresponds to two distinct regions in the PMF of the TSS-Pro-generated trajectory (Figure 6c). This type of structural change in the CTT is also observed in the training simulation of ubiquitin (Figure 7d). It should be noted that the changes in CTT orientation observed in the TSS-Pro trajectory are physically plausible and distinct from those seen in the training simulation, further supporting TSS-Pro as an independent, and at least complementary, approach to conventional MD simulation methods.

We compared root-mean-square fluctuation (RMSF) profiles of ubiquitin from the training simulation and TSS-Pro consecutive sampling (Figure 7c). Consecutive sampling reproduced the RMSF peaks observed in the flexible loops and the CTT, showing agreement with the training simulation. The observation that CTT exhibits pronounced flexibility in both the training simulation and the TSS-Pro-generated sampling is consistent with its known role in polyubiquitin chain assembly during protein degradation[67] and in mediating protein−protein interactions.[68,69]

We applied MDSCAN clustering to characterize the structural diversity within the TSS-Pro-generated consecutive sampling of ubiquitin, resulting in six distinct clusters. Clustering quality is validated by comparing the average pairwise RMSD values within (intracluster) and between clusters (intercluster): intracluster RMSD values are substantially lower than intercluster values (Figure 8a), confirming that MDSCAN effectively grouped structurally similar conformations. The centroid structures of each cluster are

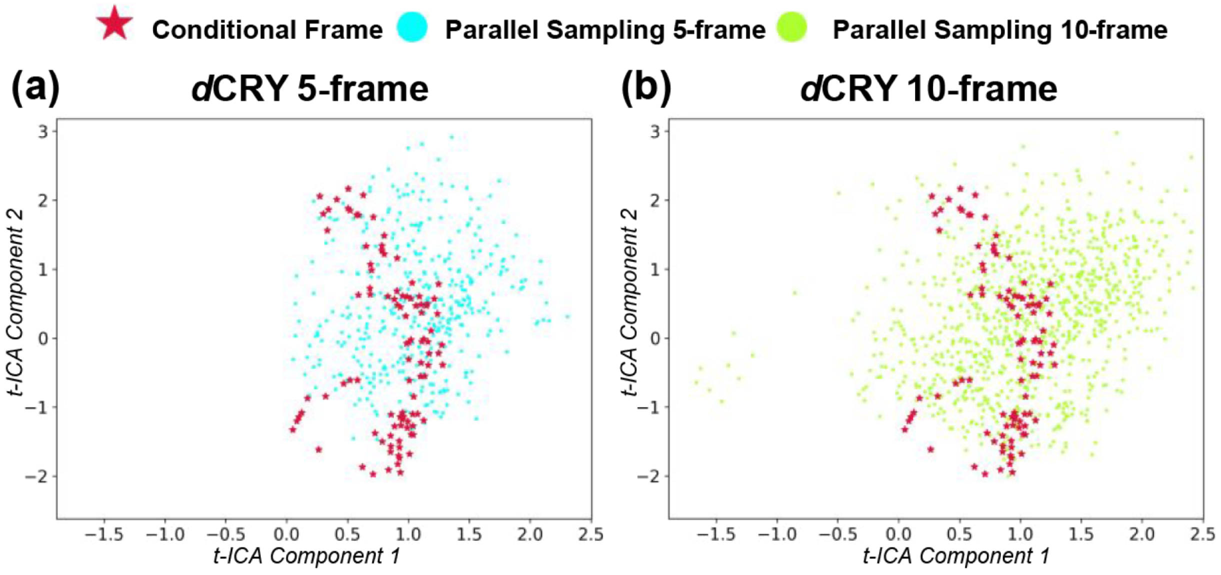


Figure 10. Parallel sampling distribution based on the conditional frames for (a) *d*CRY 5-frame and (b) *d*CRY 10-frame methods.

shown (Figure 8b). While all clusters preserve the overall ubiquitin fold, notable CTT conformational diversity is observed, highlighting the sampling efficiency of the TSS-Pro method.

We also plotted the Potential of Mean Force (PMF) distributions of ϕ−ψ angles for Ubiquitin and *d*CRY based on the TSS-Pro−generated trajectories and the 100 ns MD simulations used for model training (Figure S8). These plots provide a more detailed representation of the backbone conformational space sampled for each system. While minor under-sampling is observed in the TSS-Pro results likely due to model limitations, the overall ψ−ϕ distributions for both systems, show substantial similarity with the MD simulations. To quantify this agreement, we computed the SNR by treating the MD distribution as the signal and the deviation of the TSS-Pro distribution from it as noise. The SNR values for both Ubiquitin (4.50) and *d*CRY (3.95) are significantly above zero (Table S8), supporting the conclusion that TSS-Pro effectively captures structural features represented in the MD reference distributions.

3.5.2. Parallel Sampling Model. We also evaluated the diversity of trajectory segments generated by TSS-Pro in parallel sampling mode. To do this, we randomly selected multiple frames from the independent simulation to serve as conditioning inputs and generated multiple trajectory segments using both 5-frame and 10-frame schemes. The resulting trajectories from both schemes are visualized in the t-ICA latent space as described in Section 3.5.1 (Figure 9). The uniform distribution of these trajectory segments in the latent space demonstrates the conformational sampling efficiency of this approach. Consistent with this, the autoencoder projection shows a similar trend to the t-ICA projection (Figure S13a of the Supporting Information).

To further assess local structural flexibility, we computed RMSF values of ubiquitin residues based on the parallel sampling results. These values are elevated in the flexible loop regions and the CTT relative to other regions, and are comparable to those observed in the training simulation and the TSS-Pro-generated consecutive sampling (Figure 7c).

Overall, this proof-of-concept analysis demonstrates the potential of parallel sampling while also highlighting the need for caution when applying it for conformational sampling.

3.6. Sampling of *d*CRY as a Large Protein by TSS-Pro

Given the reduced stability observed in consecutive sampling for the large *d*CRY system, our analysis focused on the parallel sampling mode of TSS-Pro. We randomly selected frames from a 100 ns MD simulation independent of the training simulation, as indicated by red stars in Figure 10. These frames serve as conditioning inputs for generating multiple trajectory segments using both 5-frame and 10-frame schemes. The parallel sampling explores a broader region of conformational space than the conditional frames, indicating substantial structural deviations in the trajectory segments generated using TSS-Pro. The autoencoder projection shows that the parallel sampling covers a broader region of conformational space than the independent simulation (Figure S13b of the Supporting Information). Moreover, the 10-frame parallel sampling mode explores a broader region in the latent space compared to the 5-frame sampling mode. This broader coverage aligns with the higher RMSD values observed in the *d*CRY 10-frame trajectories (Figure 3c), reflecting the increased flexibility sampled over longer temporal windows. These results indicate that parallel sampling offers broad conformational coverage that scales with system size, providing improved diversity compared to consecutive sampling.

3.7. Benchmarking TSS-Pro

3.7.1. Effect of Simulation Length and Sampling Interval on Training and Generation. Given the structure of TSS-Pro's training data, consideration should be given to both the frame interval and the total simulation length used during training. Within the scope of this study, we developed four conditional generative diffusion models for ubiquitin: (1) 2 ps interval, 100 ns training set (used in the main experiments across all three model systems); (2) 2 ps interval, 1 μs training set; (3) 20 ps interval, 1 μs training set; and (4) 200 ps interval, 10 μs training set. We compared the distribution of per-trajectory average RMSD values across 50 generated samples for each model (Figure S5a). The average RMSD increases marginally (~0.2 Å) with increased frame intervals

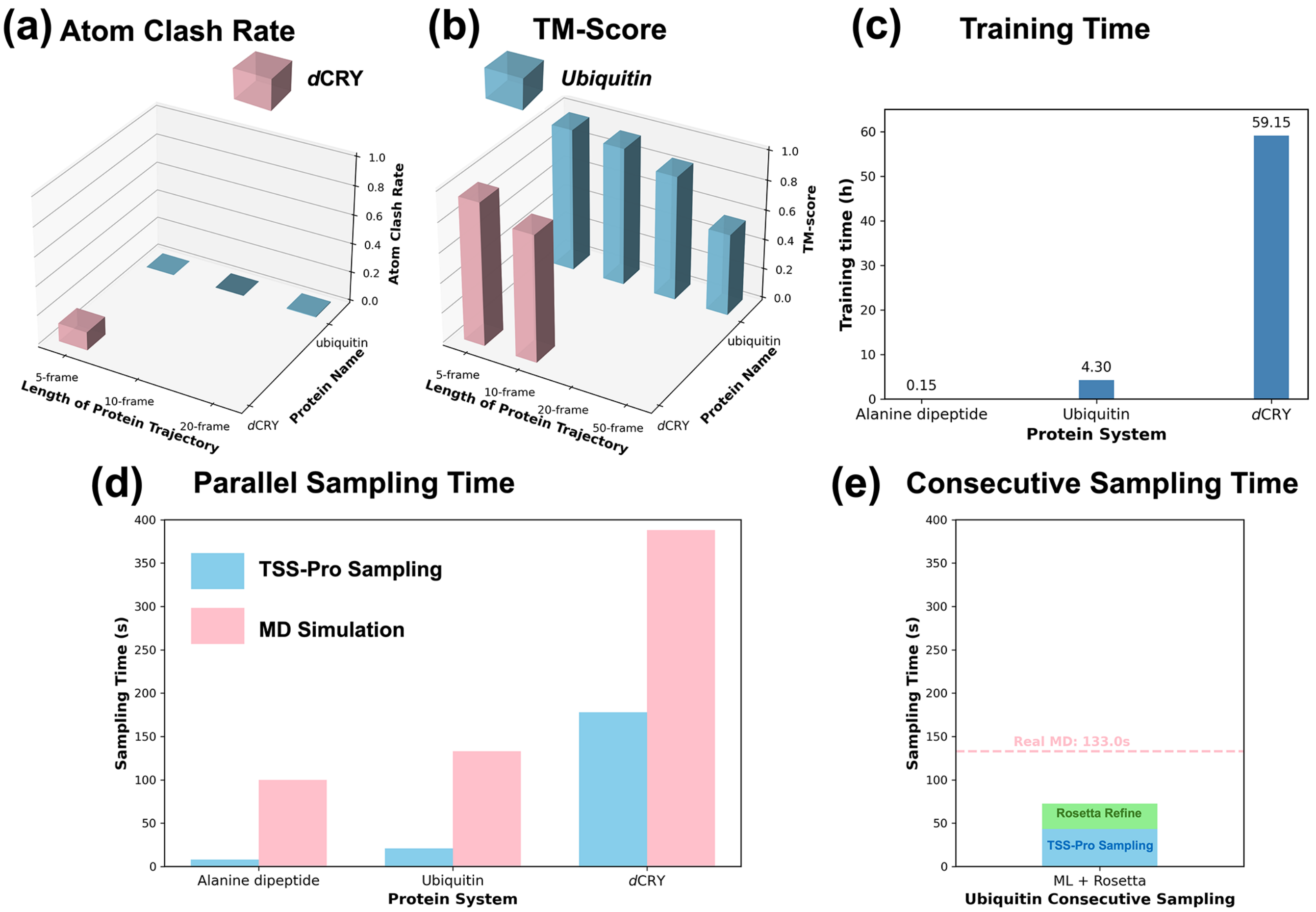


Figure 11. Benchmark of TSS-Pro assessing the quality and efficiency of protein trajectory generation. (a) Atom clash rate of generated trajectories for ubiquitin and *d*CRY, where lower values indicate more physically realistic structures. (b) TM-Score comparison between generated trajectories and reference frames, with values closer to 1.0 indicating high fidelity to native conformations and lower values indicating greater conformational diversity. (c) Training time of the TSS-Pro for the three system. Runtime benchmark showing wall-clock time required for TSS-Pro (d) parallel sampling and (e) consecutive sampling compared with conventional MD simulations across alanine dipeptide, ubiquitin, and *d*CRY.

and longer training set sizes, suggesting that these modifications have minimal impact on sampling. We further examined the RMSD relative to the first frame in the trajectory (Figure S5b) and observed that all four models exhibited similar RMSD profiles across the 10-frame prediction window. Together, these results suggest that increasing the length of MD training data or the frame interval in trajectory segments has minimal effect on the performance of TSS-Pro. However, additional analysis is required to better assess the sampling efficiency for various systems.

3.7.2. Assessment of the Accuracy and Efficiency of TSS-Pro.

To assess the quality of the generated trajectories, we calculated the atom clash rate. Steric clashes were defined as atomic pairs separated by less than 1.52 Å, corresponding to twice the carbon covalent radius (0.76 Å), a slightly more stringent threshold than the one used in prior work.[27] The clash rate is computed as the proportion of structures with at least one steric clash among all generated structures. Ubiquitin showed no steric clashes in the 5-frame and 10-frame models. However, the 20-frame segment model had a low clash rate of 0.0053 (Figure 11a). For *d*CRY, the 5-frame segment model yielded a clash rate of 0.125, which is acceptable given its size (537 residues) and the known challenges of applying transformer models to large, high-resolution proteins. By comparison, the 50-frame segment model for ubiquitin and the 10-frame segment model for *d*CRY each showed clash rates exceeding 90%, consistent with the high average RMSD values observed (Figure 3b and 3c). This suggests reduced model accuracy for larger proteins and longer trajectories. Nonetheless, the first five frames remain the most reliable and functionally relevant for the interpretation of structural features.

To evaluate structural similarity, we calculated the average TM-Score between generated trajectories and their reference frames. In both ubiquitin and *d*CRY, the average TM-Score declined as trajectory length increased (Figure 11b), consistent with the rising RMSD values observed in longer trajectories (Figure 3). This trend indicates increased conformational variability in the generated structures.

We benchmarked the efficiency of TSS-Pro sampling by measuring the time required to generate a 1 ns trajectory. For parallel sampling (Figure 11d), TSS-Pro is nearly an order of magnitude faster than conventional MD simulations for alanine dipeptide and ubiquitin. For the larger *d*CRY system, TSS-Pro is more than twice as fast as MD, highlighting its advantage in accelerating conformational sampling in large proteins. For consecutive sampling of ubiquitin (Figure 11e), TSS-Pro continues to outperform MD, even when PyRosetta optimization time is included. In contrast, training time for *d*CRY (Figure 11c) is considerably longer than for ubiquitin or

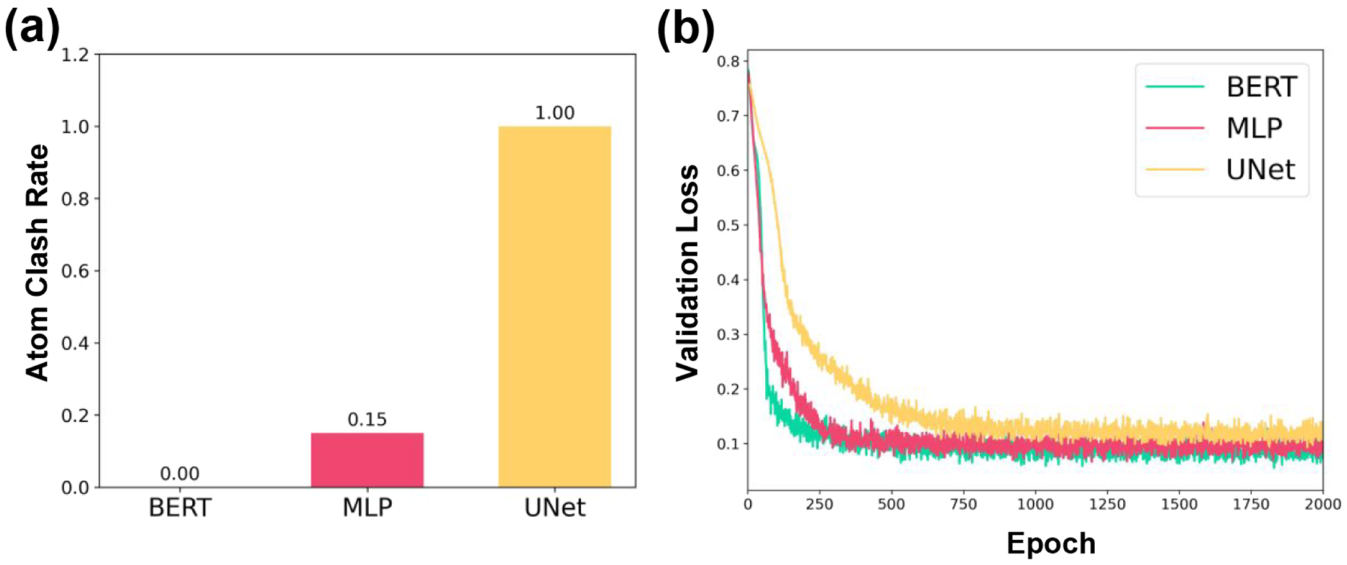


Figure 12. Comparison of the BERT, MLP, and UNet models in our cDDPM framework. (a) atomic clash rates calculated from 40 sampled structures for each model; (b) validation loss curves during the training process.

alanine dipeptide, reflecting its substantially larger system size (537 residues).

We also calculated the trajectory RMSD relative to the reference frame of the protein trajectory during the denoising process. As shown in Figure S4 of the Supporting Information, the process starts at t = 1000 as Gaussian noise, resulting in a highly random and fluctuating trajectory. As denoising progresses and the diffusion step t drops below 50, the RMSD becomes more temporally coherent; it initially increases and then stabilizes.

3.7.3. Benchmarking the BERT Model. In our cDDPM implementation, we employed a BERT-based encoder to capture long-range dependencies along the protein sequence, leveraging its attention mechanism to model long-range residue interactions in torsion-angle space. Compared with UNet and MLP architectures (see Tables S5 and S6 for model parameters of each), BERT achieves faster convergence, a lower atomic clash rate, and improved structural quality in generated trajectories (Figure 12), demonstrating its suitability for diffusion-based protein trajectory generation.

3.7.4. Integration of PyRosetta into the Consecutive Sampling Pipeline. In consecutive sampling, PyRosetta optimization was applied to preserve the folded structure of ubiquitin by rebuilding side chains and evaluating physically relevant energy terms, including van der Waals interactions, solvation, and electrostatics. A comparison of sampling results with and without PyRosetta (Figure S6 of the Supporting Information) showed that omitting optimization leads to a lower TM-Score. Rosetta energy profiles along the sampling trajectory further suggest that PyRosetta improves structural stability by maintaining sampled conformations within energetically favorable ranges.

Importantly, the integration of PyRosetta does not limit the effectiveness of TSS-Pro as a sampling method. TSS-Pro is capable of directly generating a diverse range of trajectories that can be refined using various strategies to suit specific sampling objectives. PyRosetta represents one such option, favoring energetically stable conformations. Additional refinement approaches will be developed and evaluated, including those tailored to capture rare-event trajectories that may be less energetically favorable.

Together, these analyses demonstrate that TSS-Pro generates structurally reliable and diverse trajectories with minimal steric clashes, preserves short-timescale similarity to reference structures, and offers substantial computational efficiency over conventional MD simulations.

3.7.5. Outlook and Future Work on TSS-Pro. At present, the model is limited to well-folded states of ubiquitin and is constrained by the 10 μs duration of available training data. Future work will aim to extend coverage to include unfolded and partially folded conformations to better represent the full conformational landscape. Moreover, the diffusion model operates in backbone torsion-angle space, where the denoising network is trained to predict angular noise terms rather than to directly optimize energy-related quantities. Incorporating energy-based constraints (e.g., force field or Rosetta-derived potentials) into the diffusion loss remains challenging due to the highly nonlinear relationship between local angular coordinates and global energy functions. Future work will explore hybrid approaches that incorporate coarse-grained or differentiable energy priors into the denoising process to enhance physical consistency.

4. CONCLUSIONS

Our TSS-Pro framework demonstrates strong performance in generating protein conformational trajectories with both structural fidelity and substantial computational efficiency. Evaluations on alanine dipeptide, ubiquitin, and the large *d*CRY protein show that generated trajectories exhibit low clash rates and structural similarity to reference conformations. Shorter generation segments (e.g., 5-frame model) preserve accuracy, while longer segments introduce greater conformational diversity. The method significantly accelerates sampling relative to conventional MD, achieving nearly an order-of-magnitude speedup for small and medium proteins and more than a 2-fold improvement for large systems. This capability enables efficient exploration of trajectory space and conformational landscapes, particularly for large, well-folded proteins, establishing TSS-Pro as a robust method complementing conventional MD for characterizing protein motion and conformational diversity.

■ ASSOCIATED CONTENT

Data Availability Statement

The source code for TSS-Pro is available at: https://github.com/cxiong1234/TSS_pro. The MD simulation trajectories and trained models used in this study are available in the Zenodo repository at https://zenodo.org/records/17064407.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.5c01579.

> Markov State Model (MSM) analysis; Table S1: augmented Dickey–Fuller (ADF) statistics and corresponding *p*-values for the five temporal features in alanine dipeptide, ubiquitin, and *d*CRY; Table S2: structural angle reconstruction deviation; Table S3: K–L divergence value between the MD and TSS-Pro predicted $\phi-\psi$ dihedral angle distribution for alanine dipeptide; Table S4: BERT-based conditional diffusion model parameter; Table S5: MLP-based conditional diffusion model parameter; Table S6: UNet-based conditional diffusion model parameter; Table S7: K–L divergence value between the MD and TSS-Pro predicted t-ICA distribution for ubiquitin; Table S8: Signal-to-Noise Ratio (SNR) based on $\phi-\psi$ angle distribution profiles for ubiquitin and *d*CRY; Figure S1: RMSD evolution across frames for ubiquitin trajectories generated with the 5-frame sampling under varying stochastic coefficients (ζ); Figure S2: RMSD evolution across frames for ubiquitin trajectories generated with the 10-frame sampling under varying stochastic coefficients (ζ); Figure S3: the schematic plot for the denoise process of the trajectory data $\mathrm{x}_{0:\mathrm{T}}$ as the diffusion step $t \in (0,1000)$; Figure S4: the trajectory RMSD value relative to reference frame along with the diffusion step $t \in (0,1000)$; Figure S5: (a) average RMSD values for generated trajectories of 10-frame setup; (b) RMSD as a function of frame index in the generated trajectory; Figure S6: (a) TM-Score as a function of time during consecutive sampling, with and without the PyRosetta constraint; (b) energy differences between trajectories with and without PyRosetta optimization, evaluated at different time points along the trajectory; Figure S7: schematic plot for the transition between 3D coordinates and backbone angle values; Figure S8: comparison of $\phi-\psi$ angle distributions between TSS-Pro–generated structures and 100 ns molecular dynamics (MD) simulations used for training; Figure S9: RMSD plot of the ubiquitin TSS-Pro consecutive sampling and real MD based on core region; Figure S10: the Markov State Model estimated relaxation timescale is based on the transition probabilities among different microstates using different lag times ranging from 1 to 50 steps from the trajectory of (a) 100ns independent MD simulation and (b) 100ns consecutive sampling; Figure S11: Markov State Model (MSM) identified macrostates of the (top) 100ns independent MD simulation and (bottom) 100 ns consecutive sampling; Figure S12: schematic representation of the trajectory data $x \in \mathbb{R}^{L\times(6k)}$ matrix. Each residue (rows) is described by six backbone angles (ϕ,ψ,ω,θ^a, θ^b, θ^c) across$k$ consecutive frames (columns); Figure S13: autoencoder projected conformational space of (a) Ubiquitin and (b) *d*CRY; Figure S14: Signal-to-Noise Ratio (SNR) between TSS-Pro sampling of varying trajectory lengths and MD simulations, calculated based on $\phi-\psi$ angle distributions(PDF)

■ AUTHOR INFORMATION

Corresponding Author

Peng Tao – *Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; O'Donnell Data Science and Research Computing Institute and Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States;* orcid.org/0000-0002-2488-0239; Email: ptao@smu.edu

Authors

Chuanye Xiong – *Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; O'Donnell Data Science and Research Computing Institute and Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States*

Palanisamy Kandhan – *Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; O'Donnell Data Science and Research Computing Institute and Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States;* orcid.org/0000-0002-3088-0032

Dongyang Chen – *Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; O'Donnell Data Science and Research Computing Institute and Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States;* orcid.org/0009-0000-2681-1379

Zerui Ma – *Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; O'Donnell Data Science and Research Computing Institute and Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States;* orcid.org/0009-0005-9846-6056

Eleanor D. Smith – *Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, United States; O'Donnell Data Science and Research Computing Institute and Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.5c01579

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge the computing resources provided by Southern Methodist University's O'Donnell Data Science and Research Computing Institute (ODSRCI). This work was supported by the National Institutes of Health (NIH) under Award Number R15GM122013.

REFERENCES

(1) Xia, K.; Fu, Z.; Hou, L.; Han, J.-D. J. Impacts of Protein−Protein Interaction Domains on Organism and Network Complexity. *Genome Res* **2008**, *18* (9), 1500−1508.

(2) Walhout, A. J. M.; Vidal, M. Protein Interaction Maps for Model Organisms. *Nat. Rev. Mol. Cell Biol* **2001**, *2* (1), 55−63.

(3) Taylor, G. K.; Stoddard, B. L. S. Functional and Evolutionary Relationships between Homing Endonucleases and Proteins from Their Host Organisms. *Nucleic Acids Res* **2012**, *40* (12), 5189−5200.

(4) Ivanova, V. P. Fibronectins: Structural-Functional Relationships. *J. Evol. Biochem. Physiol* **2017**, *53* (6), 450−464.

(5) Vasilchenko, A. S.; Valyshev, A. V. Pore-Forming Bacteriocins: Structural−Functional Relationships. *Arch. Microbiol* **2019**, *201* (2), 147−154.

(6) Kohen, A. Role of Dynamics in Enzyme Catalysis: Substantial versus Semantic Controversies. *Acc. Chem. Res* **2015**, *48* (2), 466−473.

(7) Cabal-Hierro, L.; Lazo, P. S. Signal Transduction by Tumor Necrosis Factor Receptors. *Cell. Signal* **2012**, *24* (6), 1297−1305.

(8) Saier, M. H., Jr. Molecular Phylogeny as a Basis for the Classification of Transport Proteins from Bacteria, Archaea and Eukarya. *Adv. Microb. Physiol* **1998**, *40*, 81−136.

(9) Mulligan, V. K.; Chakrabartty, A. Protein Misfolding in the Late-Onset Neurodegenerative Diseases: Common Themes and the Unique Case of Amyotrophic Lateral Sclerosis. *Proteins: Struct., Funct., Bioinf.* **2013**, *81* (8), 1285−1303.

(10) Zoltowski, B. D.; Vaidya, A. T.; Top, D.; Widom, J.; Young, M. W.; Crane, B. R. Structure of Full-Length Drosophila Cryptochrome. *Nature* **2011**, *480* (7377), 396−399.

(11) Akutsu, H. Strategies for Elucidation of the Structure and Function of the Large Membrane Protein Complex, FoF1-ATP Synthase, by Nuclear Magnetic Resonance. *Biophys. Chem* **2023**, *296*, 106988.

(12) Dregni, A. J.; Wang, H. K.; Wu, H.; Duan, P.; Jin, J.; DeGrado, W. F.; Hong, M. Inclusion of the C-Terminal Domain in the *β*-Sheet Core of Heparin-Fibrillized Three-Repeat Tau Protein Revealed by Solid-State Nuclear Magnetic Resonance Spectroscopy. *J. Am. Chem. Soc* **2021**, *143* (20), 7839−7851.

(13) Yuan, Y.; Kong, F.; Xu, H.; Zhu, A.; Yan, N.; Yan, C. Cryo-EM Structure of Human Glucose Transporter GLUT4. *Nat. Commun* **2022**, *13* (1), 2671.

(14) Li, Y.; Guo, Y.; Bröer, A.; Dai, L.; Bröer, S.; Yan, R. Cryo-EM Structure of the Human Asc-1 Transporter Complex. *Nat. Commun* **2024**, *15* (1), 3036.

(15) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267* (5612), 585−590.

(16) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev* **2016**, *116* (14), 7898−7936.

(17) Joshi, S. Y.; Deshmukh, S. A. A Review of Advancements in Coarse-Grained Molecular Dynamics Simulations. *Mol. Simul* **2021**, *47* (10−11), 786−803.

(18) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev* **2021**, *121* (16), 10142−10186.

(19) Poltavsky, I.; Tkatchenko, A. Machine Learning Force Fields: Recent Advances and Remaining Challenges. *J. Phys. Chem. Lett* **2021**, *12* (28), 6551−6564.

(20) Wang, Y.; Wang, T.; Li, S.; He, X.; Li, M.; Wang, Z.; Zheng, N.; Shao, B.; Liu, T.-Y. Enhancing Geometric Representations for Molecules with Equivariant Vector-Scalar Interactive Message Passing. *Nat. Commun* **2024**, *15* (1), 313.

(21) Wang, T.; He, X.; Li, M.; Li, Y.; Bi, R.; Wang, Y.; Cheng, C.; Shen, X.; Meng, J.; Zhang, H.; Liu, H.; Wang, Z.; Li, S.; Shao, B.; Liu, T.-Y. Ab Initio Characterization of Protein Molecular Dynamics with AI2BMD. *Nature* **2024**, *635* (8040), 1019−1027.

(22) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577* (7792), 706−710.

(23) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(24) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493−500.

(25) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123−1130.

(26) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature* **2023**, *620* (7976), 1089−1100.

(27) Lewis, S.; Hempel, T.; Jiménez-Luna, J.; Gastegger, M.; Xie, Y.; Foong, A. Y. K.; Satorras, V. G.; Abdin, O.; Veeling, B. S.; Zaporozhets, I.; Chen, Y.; Yang, S.; Foster, A. E.; Schneuing, A.; Nigam, J.; Barbero, F.; Stimper, V.; Campbell, A.; Yim, J.; Lienen, M.; Shi, Y.; Zheng, S.; Schulz, H.; Munir, U.; Sordillo, R.; Tomioka, R.; Clementi, C.; Noé, F. Scalable Emulation of Protein Equilibrium Ensembles with Generative Deep Learning. *Science* **2025**, *389*, No. eadv9817.

(28) Aviram, H. Y.; Pirchi, M.; Mazal, H.; Barak, Y.; Riven, I.; Haran, G. Direct Observation of Ultrafast Large-Scale Dynamics of an Enzyme under Turnover Conditions. *Proc. Natl. Acad. Sci. U. S. A* **2018**, *115* (13), 3243−3248.

(29) Umeda, R.; Satouh, Y.; Takemoto, M.; Nakada-Nakura, Y.; Liu, K.; Yokoyama, T.; Shirouzu, M.; Iwata, S.; Nomura, N.; Sato, K.; Ikawa, M.; Nishizawa, T.; Nureki, O. Structural Insights into Tetraspanin CD9 Function. *Nat. Commun* **2020**, *11* (1), 1606.

(30) Xu, J.; Wang, Y. Generating Multistate Conformations of P-Type ATPases with a Conditional Diffusion Model. *J. Chem. Inf. Model* **2024**, *64* (24), 9227−9239.

(31) Liu, B.; Boysen, J. G.; Unarta, I. C.; Du, X.; Li, Y.; Huang, X. Exploring Transition States of Protein Conformational Changes via Out-of-Distribution Detection in the Hyperspherical Latent Space. *Nat. Commun* **2025**, *16* (1), 349.

(32) Ingólfsson, H. I.; Neale, C.; Carpenter, T. S.; Shrestha, R.; López, C. A.; Tran, T. H.; Oppelstrup, T.; Bhatia, H.; Stanton, L. G.; Zhang, X.; Sundram, S.; Di Natale, F.; Agarwal, A.; Dharuman, G.; Kokkila Schumacher, S. I. L.; Turbyville, T.; Gulten, G.; Van, Q. N.; Goswami, D.; Jean-Francois, F.; Agamasu, C.; Chen, D.; Hettige, J. J.; Travers, T.; Sarkar, S.; Surh, M. P.; Yang, Y.; Moody, A.; Liu, S.; Van Essen, B. C.; Voter, A. F.; Ramanathan, A.; Hengartner, N. W.; Simanshu, D. K.; Stephen, A. G.; Bremer, P.-T.; Gnanakaran, S.; Glosli, J. N.; Lightstone, F. C.; McCormick, F.; Nissley, D. V.; Streitz,

F. H. Machine Learning−Driven Multiscale Modeling Reveals Lipid-Dependent Dynamics of RAS Signaling Proteins. *Proc. Natl. Acad. Sci. U. S. A* **2022**, *119* (1), No. e2113297119.

(33) Bera, P.; Mondal, J. How Good Is Generative Diffusion Model for Enhanced Sampling of Protein Conformations across Scales and in All-Atom Resolution? *J. Chem. Phys* **2025**, *163* (11), 114110.

(34) Benayad, Z.; Stirnemann, G. Hamiltonian Replica Exchange Augmented with Diffusion-Based Generative Models and Importance Sampling to Assess Biomolecular Conformational Basins and Barriers. *J. Chem. Theory Comput* **2025**, *21* (21), 10692−10704.

(35) Saharia, C.; Chan, W.; Chang, H.; Lee, C. A.; Ho, J.; Salimans, T.; Fleet, D. J.; Norouzi, M.Palette: Image-to-Image Diffusion Models*arXiv***2022**

(36) Wu, K. E.; Yang, K. K.; van den Berg, R.; Alamdari, S.; Zou, J. Y.; Lu, A. X.; Amini, A. P. Protein Structure Generation via Folding Diffusion. *Nat. Commun* **2024**, *15* (1), 1059.

(37) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *WIREs Comput. Mol. Sci* **2013**, *3* (2), 198−210.

(38) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput* **2015**, *11* (8), 3696−3713.

(39) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem* **2004**, *25* (9), 1157−1174.

(40) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem* **1993**, *97* (40), 10269−10280.

(41) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput* **2020**, *16* (1), 528−552.

(42) Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **2001**, *105* (43), 9954−9960.

(43) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys* **1977**, *23* (3), 327−341.

(44) Xiong, C.; Kandhan, P.; Zoltowski, B. D.; Tao, P. Structural Plasticity and Functional Dynamics of Pigeon Cryptochrome 4 as Avian Magnetoreceptor. *J. Mol. Biol* **2025**, *437*, 169233.

(45) Cheung, Y.-W.; Lai, K. S. Lag Order and Critical Values of the Augmented Dickey−Fuller Test. *J. Bus. Econ. Stat* **1995**, *13* (3), 277−280.

(46) Dickey, D. A.; Fuller, W. A. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *J. Am. Stat. Assoc* **1979**, *74* (366), 427−431.

(47) Hyndman, R. J. Yule-Walker Estimates for Continuous-Time Autoregressive Models. *J. Time Ser. Anal* **1993**, *14* (3), 281−296.

(48) Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference* SciPy 2010. DOI: 10.25080/Majora-92bf1922-011.

(49) Nichol, A. Q.; Dhariwal, P. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR, 2021; pp 8162−8171.

(50) Song, J.; Meng, C.; Ermon, S.Denoising Diffusion Implicit Models*arXiv***2022**

(51) Acheampong, F. A.; Nunoo-Mensah, H.; Chen, W. Transformer Models for Text-Based Emotion Detection: A Review of BERT-Based Approaches. *Artif. Intell. Rev* **2021**, *54* (8), 5789−5829.

(52) Yang, Y.; Xiong, C.; Tao, P.Angular Deviation Diffuser: A Transformer-Based Diffusion Model for Efficient Protein Conformational Ensemble Generation*bioRxiv***2025**640492

(53) Chaudhury, S.; Lyskov, S.; Gray, J. J. PyRosetta: A Script-Based Interface for Implementing Molecular Modeling Algorithms Using Rosetta. *Bioinformatics* **2010**, *26* (5), 689−691.

(54) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L. J.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput* **2017**, *13* (6), 3031−3048.

(55) Zhai, J.; Zhang, S.; Chen, J.; He, Q.Autoencoder and Its Various Variants*2018 IEEE International Conference On Systems, Man, And Cybernetics (SMC)*IEEE2018415−419

(56) Naritomi, Y.; Fuchigami, S. Slow Dynamics of a Protein Backbone in Molecular Dynamics Simulation Revealed by Time-Structure Based Independent Component Analysis. *J. Chem. Phys* **2013**, *139* (21), 215102.

(57) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput* **2015**, *11* (11), 5525−5542.

(58) González-Alemán, R.; Platero-Rochart, D.; Rodríguez-Serradet, A.; Hernández-Rodríguez, E. W.; Caballero, J.; Leclerc, F.; Montero-Cabrera, L. M. RMSD-Based HDBSCAN Clustering of Long Molecular Dynamics. *Bioinformatics* **2022**, *38* (23), 5191−5198.

(59) Campello, R. J. G. B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, Pei, J.; Tseng, V. S.; Cao, L.; Motoda, H.; Xu, G., Eds.; Springer: Berlin Heidelberg, 2013; pp. 160−172.

(60) Parsons, J.; Holmes, J. B.; Rojas, J. M.; Tsai, J.; Strauss, C. E. M. Practical Conversion from Torsion Space to Cartesian Space for in Silico Protein Synthesis. *J. Comput. Chem* **2005**, *26* (10), 1063−1068.

(61) Pouyani, M. F.; Vali, M.; Ghasemi, M. A. Lung Sound Signal Denoising Using Discrete Wavelet Transform and Artificial Neural Network. *Biomed. Signal Process. Control* **2022**, *72*, 103329.

(62) Poulinakis, K.; Drikakis, D.; Kokkinakis, I. W.; Spottswood, S. M. Machine-Learning Methods on Noisy and Sparse Data. *Mathematics* **2023**, *11* (1), 236.

(63) Kozen, D. C. Depth-First and Breadth-First Search. In *The Design and Analysis of Algorithms*, Kozen, D. C.; Springer: New York, 1992; pp. 19−24. DOI: 10.1007/978-1-4612-4400-4_4.

(64) Mironov, V.; Alexeev, Y.; Mulligan, V. K.; Fedorov, D. G. A Systematic Study of Minima in Alanine Dipeptide. *J. Comput. Chem* **2019**, *40* (2), 297−309.

(65) Šućur, Z.; Spiwok, V. Sampling Enhancement and Free Energy Prediction by the Flying Gaussian Method. *J. Chem. Theory Comput* **2016**, *12* (9), 4644−4650.

(66) Janson, G.; Valdes-Garcia, G.; Heo, L.; Feig, M. Direct Generation of Protein Conformational Ensembles via Machine Learning. *Nat. Commun* **2023**, *14* (1), 774.

(67) Hadari, T.; Warms, J. V.; Rose, I. A.; Hershko, A. A Ubiquitin C-Terminal Isopeptidase That Acts on Polyubiquitin Chains. Role in Protein Degradation. *J. Biol. Chem* **1992**, *267* (2), 719−727.

(68) Fang, Y.; Fu, D.; Shen, X.-Z. The Potential Role of Ubiquitin C-Terminal Hydrolases in Oncogenesis. *Biochim. Biophys. Acta, Rev. Cancer* **2010**, *1806* (1), 1−6.

(69) Xu, J.; Zhang, Y. How Significant Is a Protein Structure Similarity with TM-Score = 0.5? *Bioinformatics* **2010**, *26* (7), 889−895.