Assessing Generative Diffusion Models for Enhanced Sampling of Folded and Disordered Protein States Across Scales and in All-atom Resolution

Palash Bera* and Jagannath Mondal*

Tata Institute of Fundamental Research Hyderabad, Telangana 500046, India

Abstract

Molecular dynamics (MD) simulations are indispensable for studying the dynamic behavior of biomolecules, yet they are often constrained by timescale challenges that limit sampling efficiency. Generative machine learning (ML) models have recently emerged as promising tools to address these limitations. In this study, we revisit the potential of the Denoising Diffusion Probabilistic Model (DDPM) as an approach for enhancing the sampling and for generation of atomistically accurate conformational ensembles of biomolecules of diverse size, ranging from small folded proteins to large intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs). By training DDPM independently on MD simulation data of torsion angle and all-atom coordinates, we demonstrate their ability to effectively explore conformational space, including sparsely sampled regions, and to generate novel conformations. The performance of DDPM is evaluated on a diverse set of systems, including the Trp-cage mini-protein, the folded protein BPTI, the IDP α-Synuclein, and the IDR Ash1. Our results reveal that DDPM-generated ensembles accurately reproduce key structural

features such as secondary structure, radius of gyration, and contact maps, while also capturing the inherent conformational heterogeneity of IDPs. These findings highlight the potential of DDPM to enhance the efficiency and accuracy of biomolecular simulations, offering a scalable framework for gaining deeper insights into protein structure and dynamics.

Introduction

Understanding and predicting the conformational landscapes of proteins are pivotal for unraveling their functions, stability, and interactions within biological systems. While structured proteins adopt well-defined three-dimensional (3D) conformations that can be effectively characterized through experimental and computational methods, intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) lack such static structures. Instead, IDPs and IDRs exhibit vast conformational ensembles, which are critical for their diverse functional roles, including signaling, regulation, and molecular recognition.[1–4] Notably, IDPs are implicated in numerous human diseases,[5–9] such as neurodegenerative disorders and cancers, emphasizing the importance of understanding their long-timescale structural dynamics.

Despite significant advances in experimental and computational approaches, classical molecular dynamics (MD) simulations remain the gold standard for generating conformationally dynamic ensembles of biomolecules at atomic resolution. However, the extensive computational resources required for MD simulations pose a major challenge to the realistic exploration of long-timescale atomistic conformational ensembles. While specialized hardware platforms have enabled sampling of conformational landscapes,[10–12] their limited accessibility to the broader research community and the high computational demand for studying IDPs and IDRs have restricted their widespread use.

To address these limitations, various enhanced sampling methods have been developed. Techniques such as metadynamics,[13,14] umbrella sampling,[15,16] steered MD,[17,18] and temperature-

accelerated MD[19,20] introduce biases along predefined reaction coordinates to facilitate the sampling of rare events. Alternatively, adaptive sampling[21,22] utilizes order parameters to guide unbiased simulations, thereby enhancing the exploration of under-sampled regions. Another category of methods, including replica exchange MD (REMD),[23,24] accelerated MD,[25,26] and weighted-ensemble simulations[27,28] enhances sampling by modifying the system's global Hamiltonian, allowing energy barriers to be surmounted and diverse conformational states to be explored. While these methods have achieved notable success, they often rely on prior knowledge of reaction coordinates or incur significant computational costs.

Recently, machine learning-based approaches have emerged as powerful tools for enhanced sampling. Techniques such as VAMPnets[29] and time-lagged autoencoders[30] have been used to extract meaningful collective variables (CVs) from simulation data. Statistical methods, including Markov state models (MSMs)[31–33] and Hidden Markov Models (HMMs)[34–36] discretize continuous molecular trajectories into discrete states, providing insights into long-timescale dynamics. However, these approaches often require labor-intensive state-space discretization and struggle with sparse sampling of rare or inaccessible states, necessitating user-guided adaptive sampling.

In addition to traditional and machine learning-guided approaches, generative deep learning models have transformed the study of molecular dynamics and structural prediction. Autoencoder[37–39] and their variants, such as Variational Autoencoder (VAE),[40,41] have been utilized to uncover low-dimensional latent representations of complex molecular systems. They have also been used to reconstruct and generate conformation of biomolecules.[42–44] Generative Adversarial Network (GAN)[45,46] and Normalizing Flows[47–49] have shown promise in generating realistic molecular structures and conformations by learning high-dimensional probability distributions. Boltzmann Generators,[50] which combine generative modeling with statistical mechanics, directly sample equilibrium distributions, enabling efficient exploration of thermodynamic states. Large language models, including Long Short-Term Memory (LSTM)[51–54] and Transformers,[55,56] have advanced the prediction of state-to-state tran-

sitions by capturing temporal dependencies and long-range interactions. Diffusion-based models, such as the Denoising Diffusion Probabilistic Model (DDPM),[57,58] have emerged as powerful tools for generative tasks, often rivaling or surpassing other generative AI models in producing high-quality samples.[59]

Recent breakthroughs, such as AlphaFold[60,61] and RoseTTAFold,[62,63] have revolutionized protein structure prediction. However, while these methods excel at predicting static structures, they cannot generate dynamic conformational ensembles, especially for IDPs, which lack stable structures and exhibit extensive conformational heterogeneity. Generative models like GANs and DDPM have recently been applied to explore conformational ensembles of folded proteins and IDPs.[64,65] Nevertheless, most studies on IDPs rely on coarse-grained simulations, limiting their atomistic detail. Generating all-atom conformational ensembles for larger proteins, particularly IDPs, remains a significant challenge.

In this study, we investigate the potential of DDPM for enhanced sampling and conformational ensemble generation across diverse biomolecular systems. By training DDPM on torsion angle and all-atom coordinate data, we demonstrate their ability to capture key structural features, such as secondary structure, radius of gyration, and contact maps. Our analysis includes the Trp-cage mini-protein, the folded protein BPTI, the IDP α-Synuclein, and the IDR Ash1. Remarkably, the DDPM accurately learns the underlying distributions of the training data, generating thermodynamically consistent conformations closely aligned with original simulations. Furthermore, DDPM enhance sampling in sparsely populated regions of the free energy landscape, capturing it with remarkable accuracy. This study represents a significant advance in integrating generative AI with traditional MD simulations, offering a scalable and versatile framework for studying biomolecular systems.

Methods

Theoretical framework of the Denoising Diffusion Probabilistic Model (DDPM)

The DDPM is an ML-based generative model that is formulated by inspiring the principles of real diffusion processes.[57,58] After training, the DDPM can generate new samples starting from purely random noise. These generated samples closely follow the statistical probability distribution of the training data, making them physically realistic and representative of the original dataset.

Forward and Reverse Processes in DDPM. The DDPM operates through two complementary processes: (i) the forward or noising process, and (ii) the reverse or denoising process. In the forward process, noise is progressively added to the data over multiple timesteps, eventually converting it into pure Gaussian noise. Mathematically, if $Q(x_0)$ represents the real data distribution, the forward-nosing process can be described as a Markovian diffusion process with the following transition probabilities:

$$Q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

Here, t is the timestep, $t \in [0, T]$, T denotes the total number of timesteps, x_t represents the noisy data at timestep t, and β_t is the predefined noise schedule which controls the amount of noise added between timesteps $t-1$ and t. Using the reparameterization trick, we can rewrite equation 1 in a closed form as:

$$Q\left(x_t \mid x_0\right) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} x_0, \left(1-\bar{\alpha}_t\right) I\right) \tag{2}$$

where $\alpha_t = 1-\beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$. This reparameterization trick helps us to sample at any arbitrary timestep t (see subsection *Brief overview of reparameterization scheme in DDPM*

in the supplementary information (SI) for details).

On the other hand, the reverse process uses a neural network trained to model the reverse of the noising process by predicting the noise added at each step, effectively reconstructing the original data from the noisy input. The reverse denoising process is mathematically defined as:

$$P_\theta\left(x_{t-1} \mid x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \tag{3}$$

where μ_θ and Σ_θ represent the learnable mean and variance of the conditional distribution, parameterized by the neural network.

Model Architecture and Hyperparameters

In this work, we employ a commonly used architecture for this task, namely U-Net, an encoder-decoder neural network originally introduced by Ronneberger et al.[66] U-Net also has been widely used for tasks such as medical image segmentation.

Figure 1 represents a schematic overview of the noising and denoising processes. In our study, the input data at timestep $t = 0$ consist of either the backbone torsion angles (ϕ and ψ) or the coordinates of all heavy atoms in the proteins (excluding hydrogen atoms). We set the total timesteps for the diffusion process to $T = 1000$. The lower panel of the figure 1 illustrates the U-Net model used in our study, particularly when the input data are backbone torsion angles. The U-Net architecture primarily involves downsampling and upsampling operations. Each downsampling stage comprises ResNet blocks combined with a downsampling operation (max-pooling). Similarly, each upsampling stage consists of ResNet blocks with upsampling operations (transposed convolutions). When using heavy-atom coordinates as input data, we observed that the model required a greater number of downsampling and upsampling blocks compared to backbone torsion angles to generate realistic protein conformations. This is likely due to the significantly larger feature space represented by

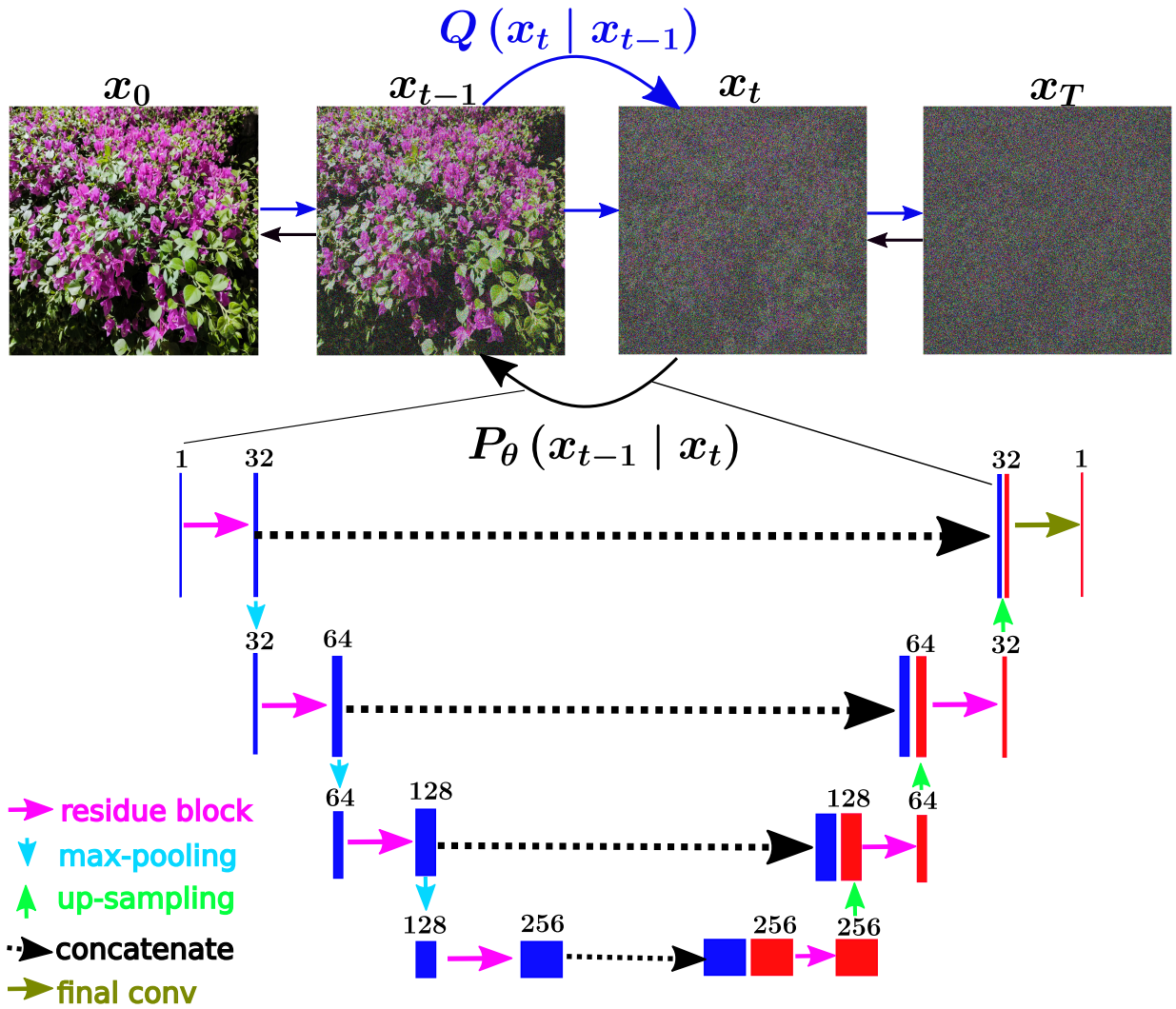

Figure 1: **A schematic overview of the noising and denoising processes.** We illustrated the noising-denoising processes using an image of a Paperflower clicked at the TIFRH campus during summer. The upper panel depicts the forward (noising) process, represented by the blue arrow, where the original input image x_0 is progressively corrupted by adding random noise, and the reverse (denoising) process, represented by the black arrow, where a neural network predicts the added noise. The lower panel illustrates the U-Net model architecture. The U-Net consists of downsampling and upsampling operations, where each downsampling stage features residue or ResNet blocks with max-pooling, and similarly each upsampling stage includes residue or ResNet blocks with transposed convolutions.

the coordinates, requiring additional layers for the diffusion model to accurately learn the underlying distribution of the training data. A comprehensive description of the U-Net architectures used for the various systems in our study is provided in the SI (see subsection

Adapting U-Net Architectures Based on Input Feature Complexity in the SI). To optimize the neural network parameters (weights and biases), we employed the Adam optimizer[67] with a learning rate 1×10^{-5}, minimizing the mean squared error (L_2 loss) between the predicted and added noise. Additionally, an exponential moving average (EMA) with a decay rate of 0.995 was applied to enhance the stability of the optimizing process. The code and comprehensive documentation for training the DDPM can be accessed on GitHub via the following link: https://github.com/palash892/DDPM_conformational_ensemble

Training datasets:

In our study, we employed MD simulation datasets provided by D. E. Shaw Research[10,68] for four distinct biomolecular systems, encompassing both folded proteins and IDPs. These systems are detailed below:

- **Trp-cage Mini-Protein:** For the folded mini-protein Trp-cage, we employed 100 μs-long trajectories. This globular 20-residue protein is frequently used as a benchmark system for computational studies.
- **α-Synuclein:** α-Synuclein, a 140-residue IDP, is characterized by its significant conformational heterogeneity. The MD simulation data spanning 73 μs were used.
- **BPTI (Bovine Pancreatic Trypsin Inhibitor):** This 58-residue folded protein was selected to test the DDPM's ability to generate all-atom conformational ensembles. We used 10 μs of unbiased MD simulation data.
- **Ash1:** The IDR Ash1 consists of 83 residues. Its conformational heterogeneity was studied using 30 μs of unbiased MD simulation data.

The percentage of data used for training the model varies across different systems and is determined based on the minima of their respective free energy surfaces along specific CVs. Since DDPM is a probabilistic model that learns the probability distribution of the training

data, this approach ensures that the data encompasses all relevant minima. However, these training data points are sparse compared to the full dataset. For example, in the case of Trp-cage, the first 10% (50,000 frames) of the total simulation data was sufficient for training, effectively capturing the essential features of the system. However, for the more complex α-Synuclein, 50% (36,562 frames) of the simulation data was necessary to achieve similar results. Notably, using only 10% (7,312 frames) of the α-Synuclein data did not work well. This limitation was attributed to insufficient frame saving in the MD simulation and the intrinsic complexity of the system, which demands a more comprehensive dataset for the model to perform effectively.

Results

Sampling Accuracy of DDPM on Benchmark Protein Systems along dihedral space

To understand the sampling accuracy of the DDPM, we first focus on two benchmark systems: the globular 20-residue mini-protein Trp-cage and the 140-residue IDP α-Synuclein. As input data for training the DDPM, we selected the backbone dihedral angles (ϕ, ψ) of these proteins. This resulted in a total of 38 torsion angles for Trp-cage and 278 for α-Synuclein. Consequently, the input data vectors are represented as: $\boldsymbol{x_0}^{\text{Trp-cage}} = [\phi_1, \psi_1, \phi_2, \psi_2 \cdots \phi_{19}, \psi_{19}]$ and $\boldsymbol{x_0}^{\alpha\text{-Synuclein}} = [\phi_1, \psi_1, \phi_2, \psi_2 \cdots \phi_{139}, \psi_{139}]$ corresponding to the Trp-cage mini-protein and α-Synuclein, respectively.

2D Free Energy Surfaces. To visualize the high dimensional torsion angles of these systems, we performed a principal component analysis (PCA) on the torsion angles and plotted the free energy surface (FES) using the first two principal components (PC-1 and PC-2). Figures 2(a) and (d) present the 2D FES plots along PC-1 and PC-2 for Trp-cage and α-Synuclein, respectively. To examine the capability of DDPM for enhanced sampling, we

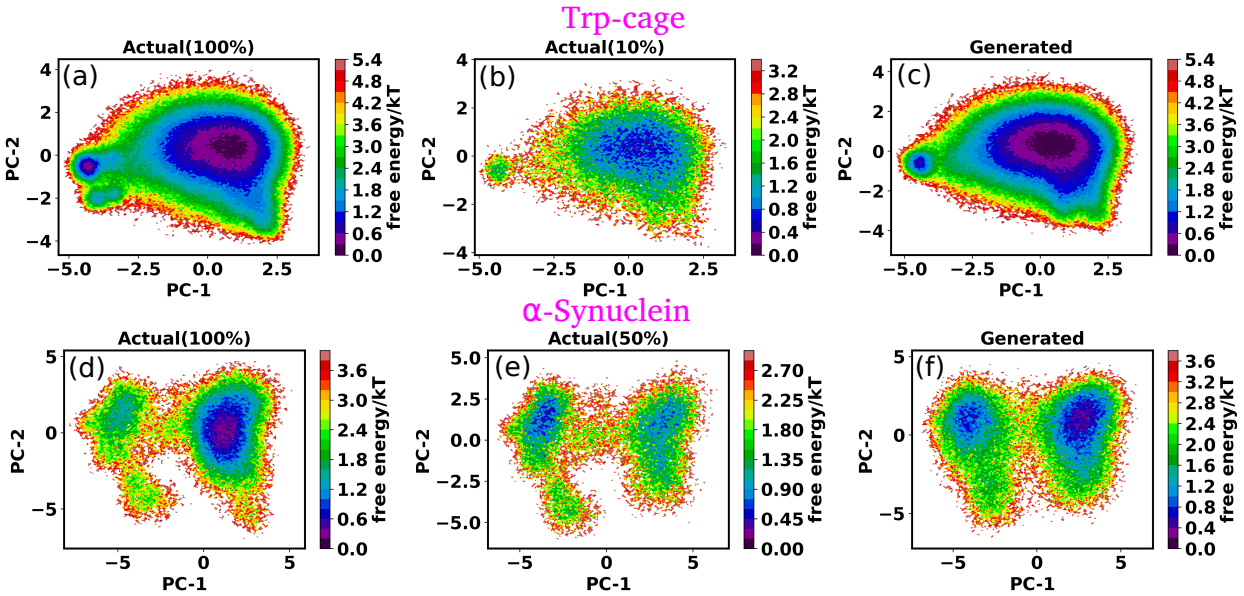

Figure 2: **The free energy surfaces (FES) of Trp-cage and α-Synuclein based on principal component analysis (PCA) of torsion angles.** (a) and (d) show the 2D FES plots along the first two principal components (PC-1 and PC-2) derived from the full molecular dynamics (MD) simulation datasets for Trp-cage and α-Synuclein, respectively. (b) and (e) depict the FES plots using subsets of the simulation data, corresponding to 10% of the total data for Trp-cage and 50% for α-Synuclein. These subsets highlight the sparsity of the sampled free energy surfaces compared to the full datasets. (c) and (f) present the FES plots generated from the DDPM-trained models, which were trained on the reduced subsets. Remarkably, the DDPM-generated FES plots closely resemble those from the full datasets and effectively sample the sparse regions observed in (b) and (e).

trained the model on subsets of the simulation data instead of the entire time series. For Trp-cage, we used the first 10% (50000 frames) of the total simulation data, while for α-Synuclein, we selected the first 50% (36562 frames) . Using these subsets, we generated 2D FES plots along PC-1 and PC-2 (Figures 2(b) and (e)). These plots indicate the sparsity in the free energy surface compared to those derived from the full 100% datasets. We then trained the DDPM on these selected subsets and generated the same number of samples as in the original 100% simulation datasets. The generated dihedral angles were projected onto the principal components defined by the original MD data. Figures 2(c) and (f) depict the corresponding FES plots for Trp-cage and α-Synuclein. Quite interestingly, the DDPM-generated FES plots closely resemble those derived from the full 100% simulation datasets. More importantly, the

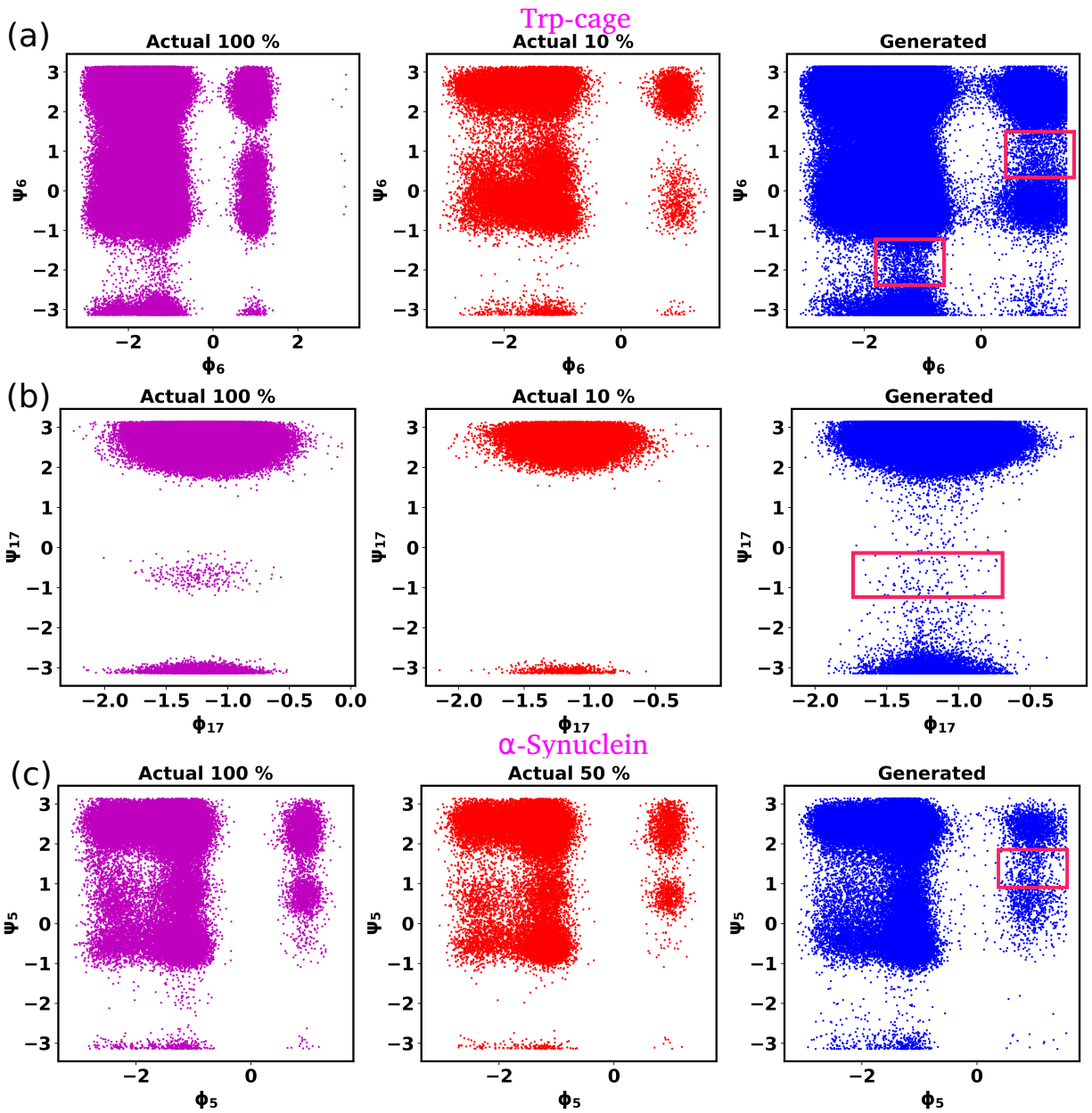

Figure 3: **Ramachandran plots illustrating the mechanisms underlying enhanced sampling.** (a) and (b) show the Ramachandran plots for the $\phi-\psi$ combinations ($\phi_6-\psi_6$) and ($\phi_{17}-\psi_{17}$) in Trp-cage, respectively. (c) presents the plot for the $\phi-\psi$ combination ($\phi_5-\psi_5$) in IDP α-Synuclein. We plotted the complete simulated data, the selected subsets for training (10% and 50%), and the DDPM-generated data. The red boxes in (a) and (c) (Generated) highlight regions with sparse sampling in the training data that are significantly enhanced in the DDPM-generated data, illustrating interpolation capabilities. In (b) (Generated), the red box highlights new low-probability samples generated by DDPM, absent in the training data but present in the full 100% simulation dataset.

sparse regions observed in Figures 2(b) and (e) were effectively sampled by the data generated by DDPM. Collectively, these results suggest that DDPM accurately learns the distribution of the training data, enabling it to generate thermodynamically consistent torsion data.

Characterizing the origin of DDPM-augmented Enhanced Sampling.

To investigate the mechanisms underlying the enhanced sampling, we analyzed the Ramachandran plots of specific $\phi - \psi$ combinations for both systems. We plotted both the full simulated data, selected subsets as previously described, and DDPM-generated data. Figures 3(a) and (b) show the representative Ramachandran plots of $(\phi_6 - \psi_6)$ and $(\phi_{17} - \psi_{17})$ for Trp-cage mini-protein, while Figure 3(c) depicts the plot of $(\phi_5 - \psi_5)$ for IDP α-Synuclein. Additional combinations of $\phi - \psi$ for both systems are provided in the Supporting Information (SI) (Figures S1 and S2). For α-Synuclein, which has 139 possible consecutive $\phi - \psi$ combinations, the first 5 combinations are presented in the SI. Our analysis reveals two key mechanisms for enhanced sampling: (i) The DDPM can effectively enhance sampling in regions where training data is sparsely distributed, as highlighted by the red boxes in Figures 3(a) and (c). These regions, which are sparsely populated in the training data, are significantly better sampled in the DDPM-generated data. (ii) The DDPM can generate new samples with very low probabilities, absent in the training data but present in the full 100% simulation dataset. This mechanism is illustrated in Figure 3(b), where the highlighted region in the DDPM-generated data corresponds to previously unsampled areas of the Ramachandran plot. These observations are also consistent with the previous study by Wang et al.[69] Nevertheless, our findings indicate DDPM's ability to enhance sampling through both interpolations of existing data distributions and generation of previously unobserved samples.

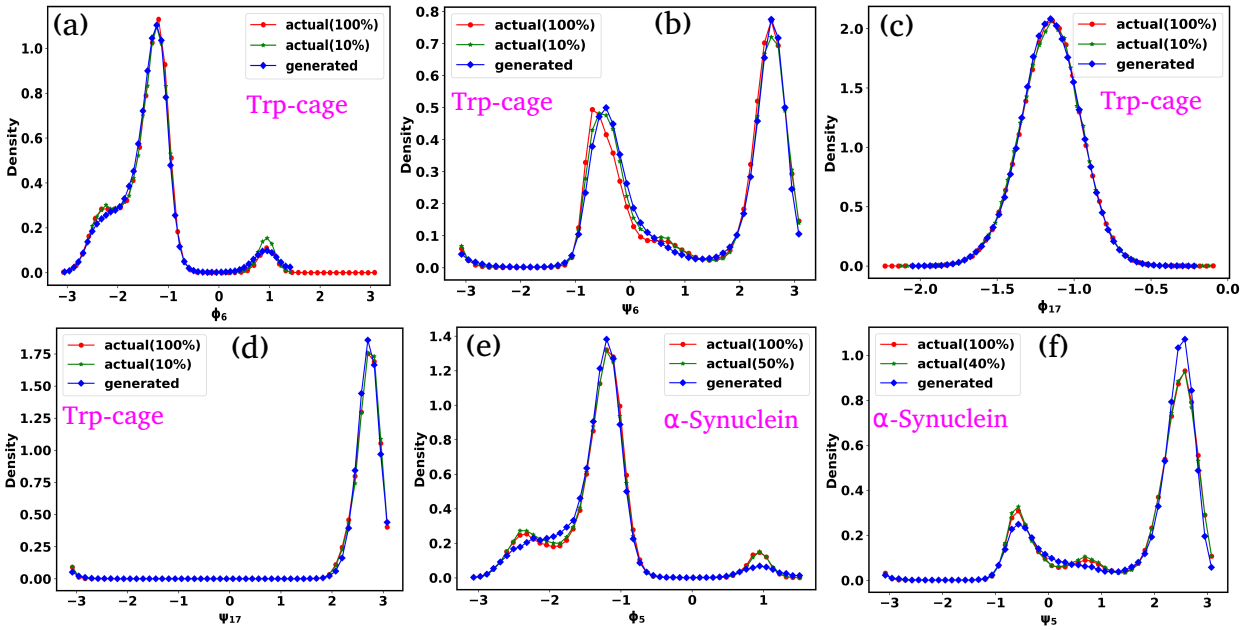

Figure 4: **Comparison of individual probability distributions for selected torsion angles ϕ and ψ.** (a-d) Probability distributions for selected dihedral angles in Trp-cage mini-protein. (e-f) Probability distributions for selected dihedral angles in α-Synuclein. The DDPM-generated distributions closely match the original simulated data. More importantly, the DDPM accurately captures the bimodal distribution of torsion angles, ((b) and (f)).

Probability Distributions of Torsion Angles.

To further validate DDPM's performance, we compared the individual probability distributions of torsion angles ϕ and ψ. For this comparison, we chose the complete simulated data, a subset of simulated data, and DDPM-generated data, as previously outlined. Figures 4(a-d) depict the probability distributions of selected dihedral angles for Trp-cage mini-protein, while Figures 4(e-f) illustrate these distributions for α-Synuclein. Additional distributions for other ϕ and ψ angles are reported in the SI (Figures S3 and S4). The results indicate that for both systems, the distribution of generated torsion angles closely aligns with that of the original data. Notably, the DDPM captures the bimodal distribution of torsion angles quite accurately, as demonstrated in Figures 4(b) and (f). This accuracy highlights DDPM's ability to learn and generate complex distributions within molecular dynamics simulations.

Training DDPM on All-Atom Protein Coordinates

In the preceding sections, we discussed the ability of the DDPM to explore sparsely sampled and previously unsampled regions of the free energy landscape, highlighting its potential as a powerful tool for enhancing sampling in molecular dynamics simulations. However, our previous analyses used backbone torsion angles as input data for training the DDPM, which restricted its output to torsion angles alone. While torsion angles are insightful, they can not fully describe the structural properties of complex biomolecules. To address this limitation, we expanded our approach by training the DDPM using the coordinates of all heavy atoms as input data. This approach enables the generation of complete conformational ensembles of biomolecules, from which various structural properties can be calculated. In addition to the Trp-cage mini protein and the IDP α-Synuclein, we included two additional systems: the 58-residue folded protein BPTI and the 83-residue IDR transcription factor Ash1. Specifically, we selected 40% (19,960 frames) of the BPTI trajectory and 25% (42,500 frames) of the Ash1 trajectory to train the DDPM. For Trp-cage and α-Synuclein, we used the same fraction of the initial MD simulation data as in previous analyses. All the aforementioned percentages of the trajectories are selected from the initial portion of the MD simulation data. For DDPM training, we first excluded hydrogen atoms from the trajectories, then centered them at the origin and aligned them with the first frame of each respective trajectory.[44,70] Subsequently, these aligned coordinates were scaled between -1 and +1 to serve as input data for the DDPM. After generating coordinates with the DDPM, these outputs were scaled back to their real-space values across all frames.

Evaluation of Conformational Ensembles for Folded Proteins

Validation Using Trp-cage Mini Protein. To evaluate the structural accuracy of the conformational ensembles generated by the DDPM, we focused initially on two folded proteins: Trp-cage and BPTI. For the Trp-cage mini protein, we calculated backbone torsion angles from the generated conformations and performed PCA. Figures S5 show FES plots

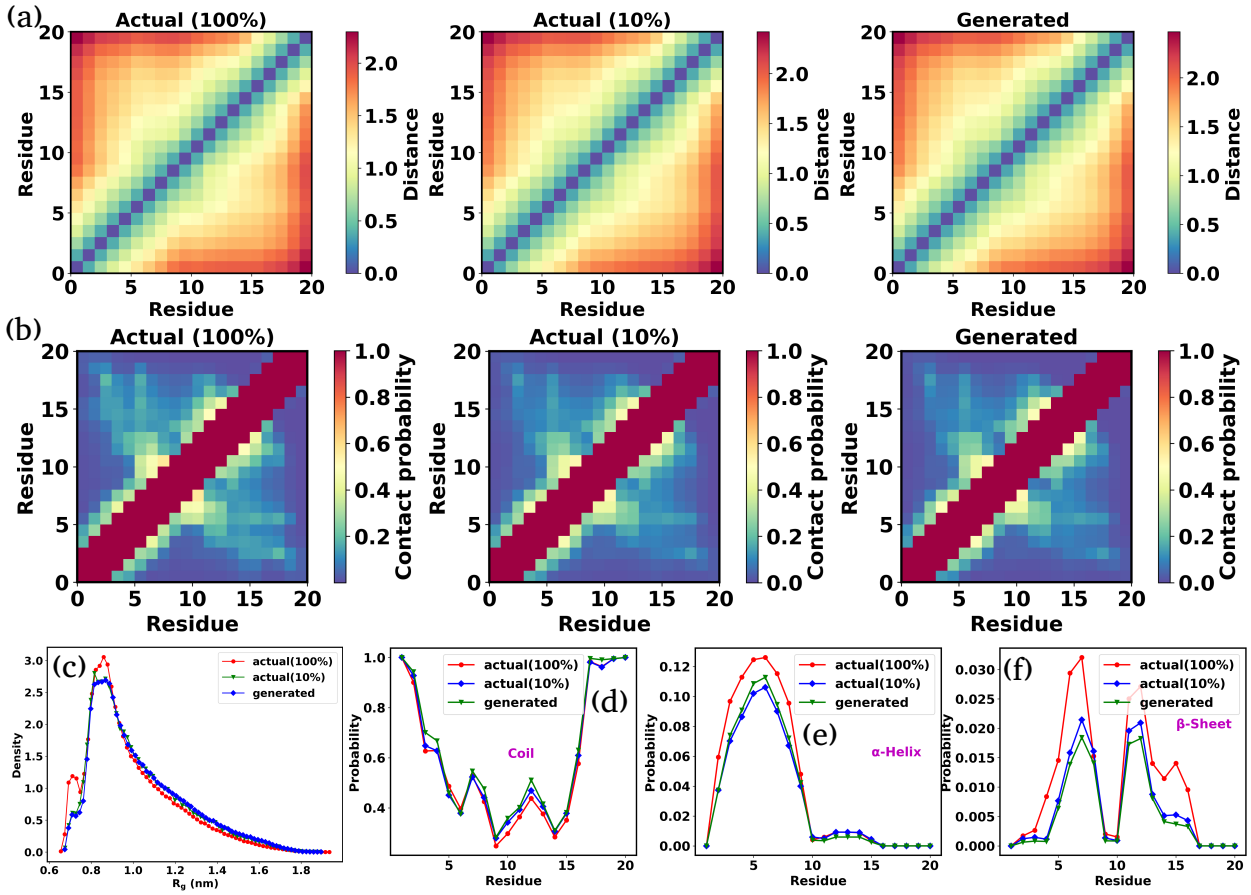

Figure 5: **Structural analysis of DDPM-generated conformations for the Trp-cage mini-protein.** (a) Distance map and (b) contact map showing pairwise distances and contact probabilities between $C\alpha$ atoms, calculated from the original 100% MD simulation data, the training subset, and DDPM-generated data. (c) Distribution of radius of gyration (R_g) values for the complete 100% MD simulation data, training subset, and DDPM-generated data. (d-f) Secondary structure probability distribution as a function of residue index for Coil, α-Helix, and β-Sheet, respectively, for the original, training subset, and DDPM-generated conformations. The results suggest that the DDPM-generated ensembles accurately replicate the structural features of the original data, preserving both global compactness and local secondary structure characteristics.

of the torsion angles along the two principal axes, PC-1 and PC-2. These plots reveal that the FES derived from the DDPM-generated data aligns closely with the original 100% MD simulation data. Furthermore, the generated conformations enhanced sampling in regions sparsely sampled in the training data (10%). These observations are consistent with our previous analyses of FES plots obtained when training the DDPM directly on torsion angle data, reaffirming the model's ability to enhance sampling effectively.

Next, we analyzed additional structural properties, such as the average pairwise distance and contact probability between residues ($C\alpha - C\alpha$). Pairwise distances were calculated for $C\alpha$ atoms across all frames, and contact probabilities were determined by defining a threshold of 0.8 nm. A contact between two $C\alpha$ atoms was assigned a value of 1 if their distance was below the threshold, and 0 otherwise. These calculations were performed for each frame and then averaged across all frames. Panels (a) and (b) in Figure 5 represent the distance map and contact map for the Trp-cage mini protein, derived from the original 100% MD simulation data, the subset of training data, and the DDPM-generated data. These maps demonstrate that the DDPM-generated conformations accurately capture the structural features of the protein, showing excellent agreement with the original data while improving sampling beyond the limited training set.

To evaluate compactness and local structural features, we compared the distribution of radius of gyration (R_g) and the secondary structure between the actual and DDPM-generated conformations. We calculated the counts of secondary structures, mainly Coil, α-Helix, and β-Sheet, for each residue in every frame and converted these counts into probabilities by dividing them by the total number of frames. Figure 5(c) represents the distribution of R_g values for the complete 100% MD simulation data, the training subset, and the DDPM-generated data. Similarly, Figures 5(d-f) show the secondary structure probabilities as functions of residue indices for Coil, α-Helix, and β-Sheet, respectively. The results indicate that the DDPM-generated ensembles closely replicate the conformational features of the original simulation data, maintaining both global compactness and local secondary structure characteristics with high accuracy.

Validation Using Folded Protein BPTI.

In a similar spirit, we also generated the conformational ensemble for the larger 58-residue folded protein BPTI. Backbone torsion angles were calculated from the generated as well as original conformations, followed by PCA to extract principal components. Figure 6(a)

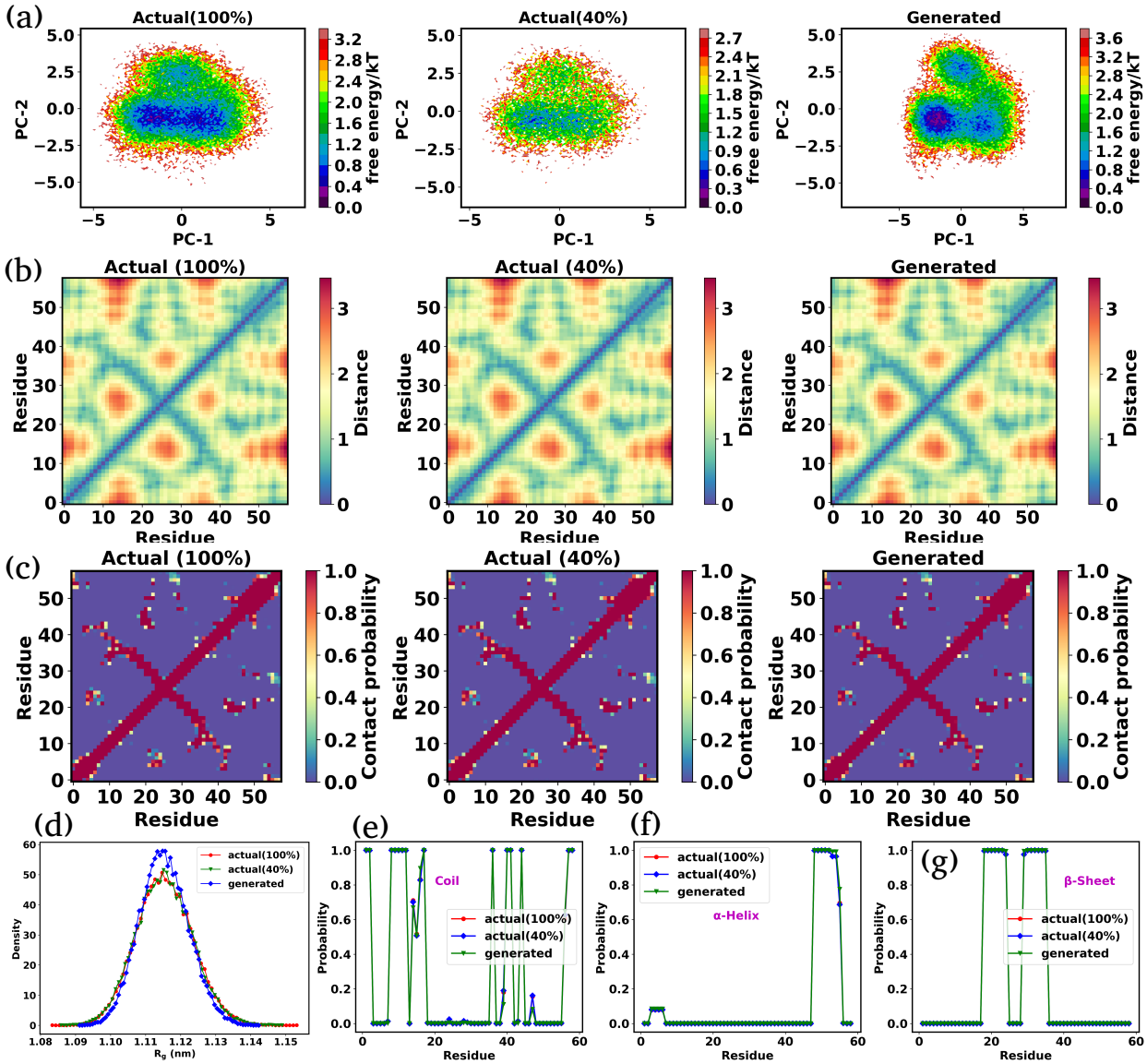

Figure 6: **Structural analysis of the DDPM-generated conformational ensemble of folded protein BPTI .** (a) FES plots along the first two principal components (PC-1 and PC-2) for the full 100% MD simulation data, the training subset, and the DDPM-generated data, demonstrating that the DDPM-generated all-atom conformations accurately replicate the trends in the FES while effectively sampling sparsely populated regions. (b) Distance map, (c) contact map, and (d) (R_g) distributions, highlighting the structural similarities between the DDPM-generated and original data. (e-g) Secondary structure probabilities for Coil, α-Helix, and β-Sheet respectively.

presents the FES plots along the first two principal components (PC-1 and PC-2) for the complete 100% MD simulation data, the training subset, and the DDPM-generated data. The generated conformations reproduce trends in the FES that align closely with the original

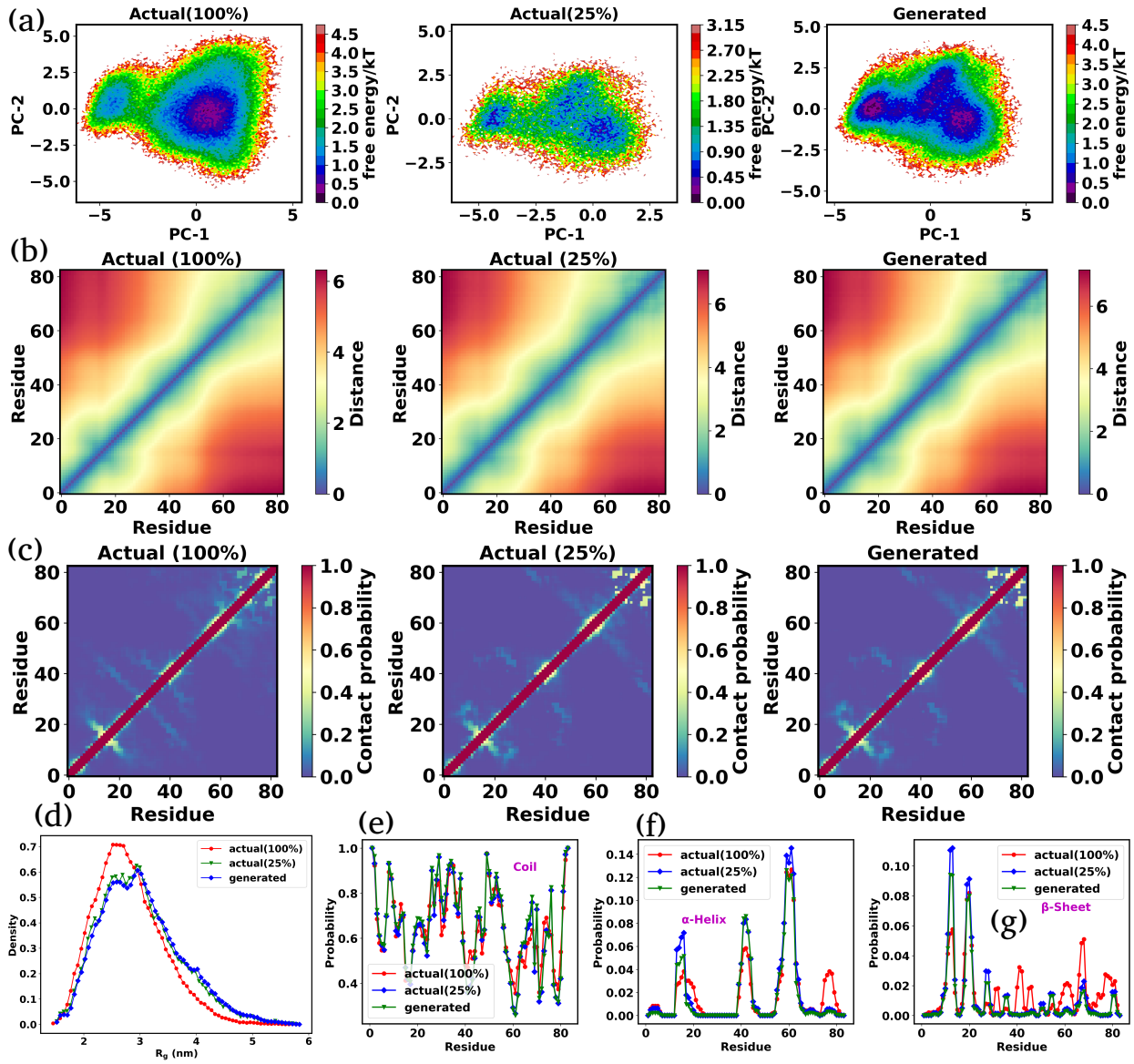

Figure 7: **Structural analysis of the intrinsically disordered region (IDR) Ash1** (a) 2D free energy surface (FES) plots along the first two principal components (PC-1 and PC-2) derived from backbone torsion angles. (b) Distance map, and (c) contact map showing the average pairwise $C\alpha - C\alpha$ distances and contact probability across all frames. (d) Distribution of the radius of gyration (R_g) for the full dataset, training subset, and DDPM-generated data. (e-g) Secondary structure probabilities for Coil, α-Helix, and β-Sheet as functions of residue indices. The plots demonstrate that the DDPM-generated conformations closely reproduce the structural properties of the original data, capturing both global and local features with high accuracy. Slight deviations in absolute secondary structure probabilities are observed for a few residues, but overall trends remain consistent.

full dataset, while also sampling sparsely populated regions of the training data effectively. Additional comparisons of the distance map, contact map, R_g distribution, and secondary structure (Figures 6(b), (c), (d), and (e-g), respectively) further validate that the DDPM-generated conformations closely match the structural properties of the original data. Notably, the secondary structure plots indicate a probability of 1 for certain residues, confirming that BPTI maintains a rigid and consistent secondary structure

Evaluation of Conformational Ensembles for IDP and IDR

Validation Using IDR Ash1.

In the earlier section, we highlighted the capability of DDPM in accurately generating conformations of folded proteins, specifically the 20-residue globular protein Trp-cage and the 58-residue folded protein BPTI. Now, we shift our focus to intrinsically disordered systems, IDRs and IDPs, which lack stable configurations and instead form dynamic and heterogeneous ensembles of conformations. We first examined the 83-residue IDR Ash1 using all-atom molecular dynamics simulations. In a similar fashion to folded protein, backbone torsion angles were calculated from the conformations, and PCA was performed on the torsion angle data. Figure 7(a) shows the 2D FES plots along the first two principal components (PC-1 and PC-2) for the complete MD simulation data, the training subset, and the DDPM-generated data. These plots suggest that the FES plot by the DDPM-generated conformational ensemble of the IDR also very nicely captured the FES plot of the original data and the sparsely sampled region in the training data is also well sampled by the DDPM-generated data.

Similarly, we analyzed the distance map (Figure 7(b)), contact map (Figure 7(c)), radius of gyration distribution (Figure 7(d)), and secondary structure probabilities (Figures 7(e-g)) for the IDR Ash1. These analyses reveal that the DDPM-generated conformational ensemble aligns closely with the structural properties of the original 100% MD simulation data. The distance and contact maps demonstrate excellent agreement, capturing both local and global structural features. The R_g distribution also closely replicates the compactness and

heterogeneity observed in the original dataset. While the secondary structure probabilities for Coil, α-Helix, and β-Sheet formations show trends similar to the original data, some residues exhibit slight deviations in absolute probabilities, particularly for α-Helix and β-Sheet. However, the overall trends remain consistent, confirming the ability of the DDPM to capture the conformational heterogeneity and dynamic structural features characteristic of IDRs.

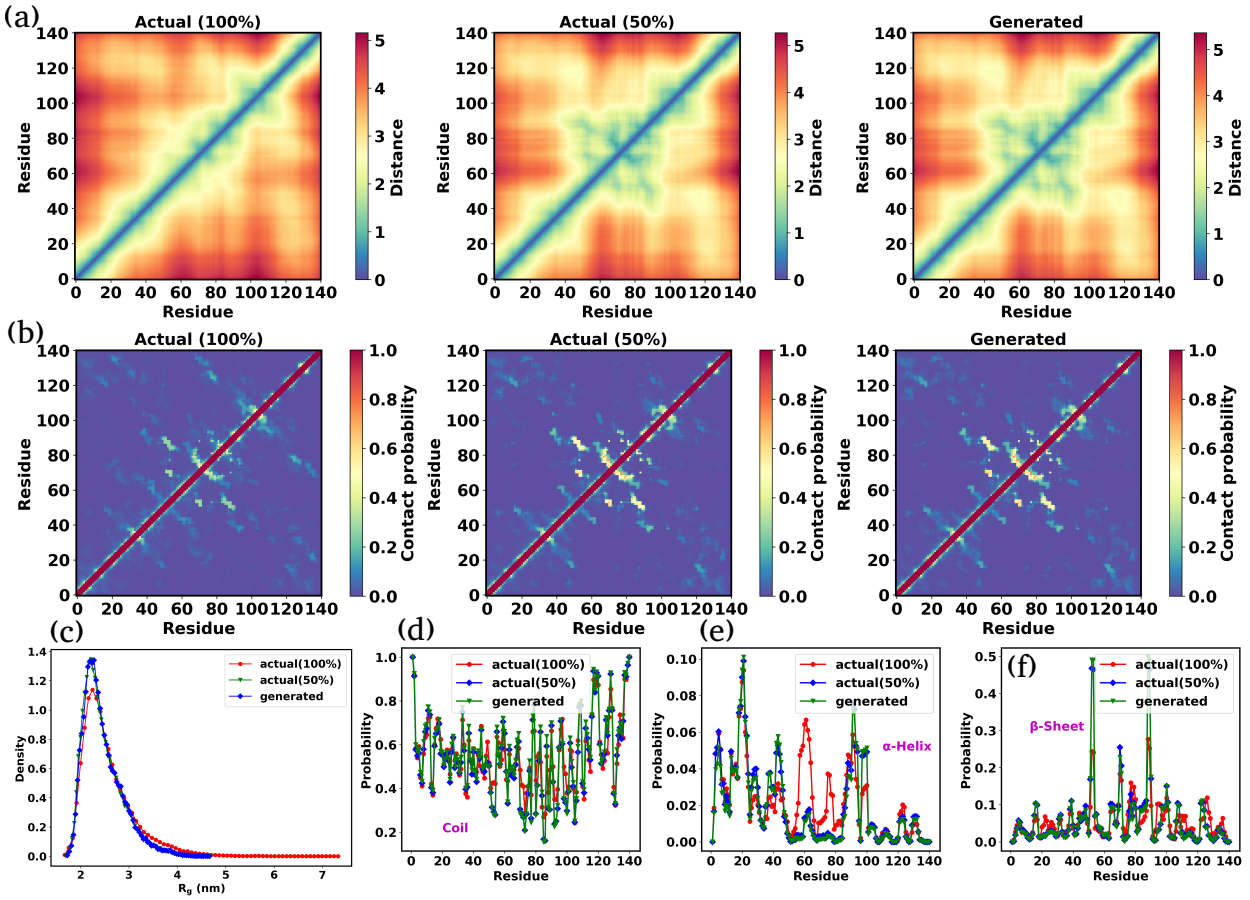

Figure 8: **Comparison of structural features for α-Synuclein obtained from MD simulation, training data, and DDPM-generated ensembles.** (a) Distance map and (b) contact map comparing the DDPM-generated conformations with the original MD data, showing excellent agreement in capturing local and global structural features. (c) The (R_g) distribution demonstrates that the DDPM-generated conformations closely align with the compactness and heterogeneity of the original dataset. (df) Secondary structure probabilities for Coil, α-Helix, and β-Sheet formations, highlighting trends consistent with the original data, with slight deviations observed in α-Helix probabilities for certain residues. These results confirm that DDPM accurately captures the conformational heterogeneity and structural properties of α-Synuclein..

Validation Using IDP α-Synuclein.

To further test the capability of DDPM in generating conformational ensembles, we investigated whether the model could accurately replicate the structural properties of a larger IDP. Specifically, we applied DDPM to α-Synuclein, a 140-residue IDP known for its highly dynamic and heterogeneous ensemble of conformations. Figure S6 presents the 2D FES plots along the first two principal components (PC-1 and PC-2) of the backbone torsion angle for the full MD simulation data, the training subset, and the DDPM-generated data. Remarkably, for this larger and more complex IDP, the DDPM-generated conformations effectively recapitulate the essential features of the original FES while also sampling regions sparsely populated in the training data, showcasing the model's ability to capture the full conformational landscape. Furthermore, the generated FES landscape is also consistent with our previous analyses of FES plots obtained by training the DDPM directly on torsion angle data.

Additional analyses were performed to validate the structural properties of α-Synuclein. These include the distance map (Figure 8(a)), contact map (Figure 8(b)), radius of gyration (R_g) distribution (Figure 8(c)), and secondary structure probabilities (Figures 8(d-f)) for Coil, α-Helix, and β-Sheet formations. The results reveal excellent agreement between the DDPM-generated conformations and the original 100% MD simulation data. While slight deviations in absolute probabilities were observed for α-Helix formation in certain residues, the overall trends are consistent with the original data. These analyses confirm that the DDPM-generated conformations accurately represent the structural heterogeneity of α-Synuclein, including its local and global structural properties.

Structural Comparison of DDPM-Generated Configurations with Original Data

To visualize how closely the structure of the generated proteins aligns with the original data, we individually plotted the configurations. First, we extracted the configurations of each protein from their respective minima on the FES plots of the original dataset and used them as reference structures for calculating root-mean-square deviations (RMSDs). We computed the RMSDs of the DDPM-generated ensemble relative to these reference structures and selected five configurations for each protein with the smallest RMSD values. Subsequently, we overlaid these five DDPM-generated configurations with their respective reference structures from the original simulation data. In all overlay figures, the reference structures are shown in blue, while the generated structures are in red

Figures 9(a) and (b) show the superimposed configurations of the Trp-cage mini-protein, revealing its folded α-helix and unfolded random coils, respectively. Similarly, Figures 9(c-e) visualize the superimposed configurations of the folded BPTI protein. For both Trp-cage and BPTI, the DDPM-generated configurations closely match the secondary structure observed in the reference data. Figures 9(f-g) and (h-j) present the overlaid structures of the IDR Ash1 and IDP α-Synuclein, respectively. These visualizations confirm that the DDPM-generated configurations not only closely replicate the original IDP structures but also effectively capture their significant conformational heterogeneity.

Discussions

The findings presented in our study demonstrate the efficacy of the DDPM in enhancing sampling accuracy within MD simulations, ranging from small folded proteins to large IDPs. By using the principles of diffusion processes, DDPM not only generates realistic molecular conformations but also effectively samples regions of the conformational space that are sparsely populated in the training data. This capability is particularly significant when considering

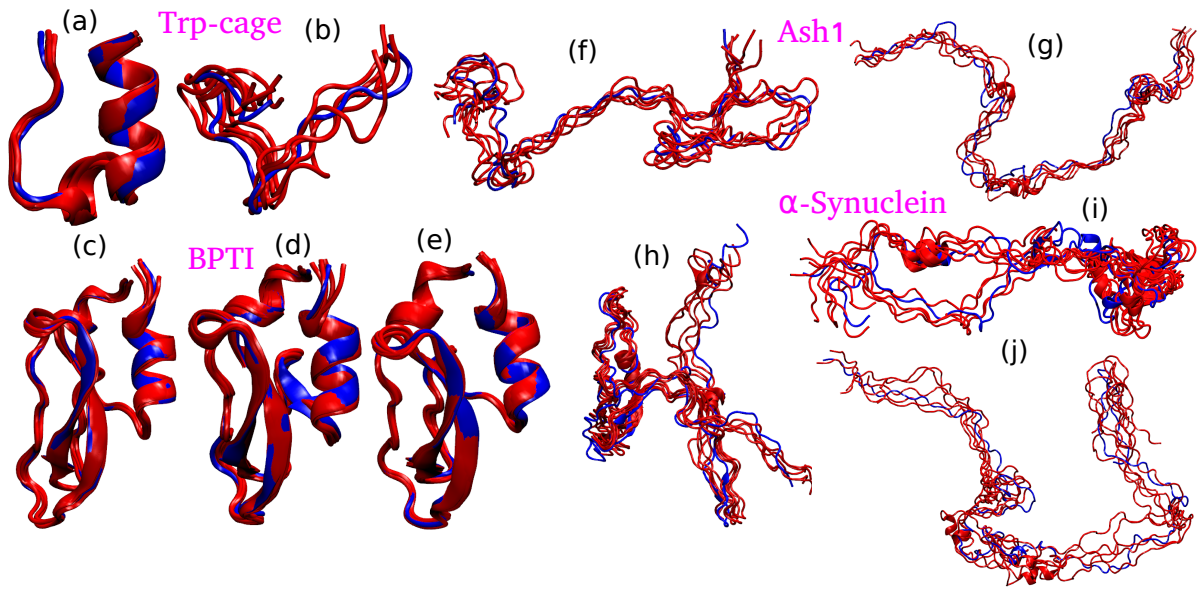

Figure 9: **Overlay of DDPM-generated configurations (red) with reference structures (blue) extracted from the minima of FES plots of the original dataset.** (a-b) Superimposed configurations of the Trp-cage mini-protein, illustrating its folded α-helix and unfolded random coil states, respectively. (c-e) Overlay of the folded BPTI protein configurations, showcasing the accurate reproduction of secondary structures by the DDPM-generated ensembles. (f-g) and (h-j) Superimposed configurations of IDR Ash1 and IDP α-Synuclein, respectively, demonstrating the ability of DDPM to closely capture the conformational heterogeneity of IDPs.

the challenges associated with exploring high-dimensional biomolecular landscapes.

Our study began with training DDPM on torsion angle for Trpcage mini-protein and IDP α-Synuclein. Remarkably, the FES generated by DDPM closely resembled those obtained from full-length MD simulations, even when trained on significantly smaller subsets of data. The analysis of torsion angle distributions revealed that DDPM accurately captured complex features, including bimodal distributions. Our findings highlight two primary mechanisms through which DDPM enhances sampling: (i) interpolation within existing data distributions and (ii) the generation of new, low-probability samples.

Expanding upon this, we trained DDPM using all-atom coordinates and observed its remarkable ability to comprehensively explore the conformational spaces of both folded proteins and IDPs. Through comparisons across a range of structural properties including free energy surfaces, inter-residue distances, contact maps, radius of gyration distributions, and

secondary structure probabilities, DDPM-generated conformational ensembles demonstrated perfect alignment with the original MD simulation data. These results affirm the model's ability to generate conformations that are not only statistically representative but also structurally relevant.

While our study demonstrates the significant potential of DDPM in mining conformational ensembles for biomolecular simulations, there are some limitations to consider. One key limitation is that the generated conformations in the ensemble are temporally discontinuous resulting in a loss of kinetic information for the respective systems. Additionally, the DDPM used in this study is not transferable, meaning that a separate model must be trained for each protein to generate its conformational ensemble. Recent advancements in modeling IDPs have introduced transferable generative models capable of generating conformational ensembles for proteins based on their residue sequences.[64,65] However, these approaches are primarily limited to coarse-grained models of IDPs. For atomistic systems, achieving transferability would require equilibrated training data for numerous proteins, which is computationally expensive, especially for large and complex IDPs. Future work could focus on developing transferable atomistic DDPM by training on an extensive dataset of all-atom proteins, though this would necessitate substantial computational resources.

Our study primarily considers simulated protein data under specific physical conditions, such as fixed temperature, pressure, etc., resulting in conformational ensembles representative of those particular conditions. In the future, extending the model to generate conformational ensembles across a range of physiological conditions by training it under certain other conditions would be a valuable advancement.[69,71] Moreover, while this work focuses exclusively on protein conformational ensembles, future research could explore applying DDPM to more complex systems, including protein-ligand binding,[72–75] membrane-associated IDPs,[76,77] and IDP aggregation.[78–80] These extensions would significantly broaden the scope of DDPM applications in biomolecular simulations.

Data and code availability

All data supporting the findings are included within the manuscript. Additionally, the code and comprehensive documentation for training the DDPM can be publicly accessed on GitHub via the following link: https://github.com/palash892/DDPM_conformational_ensemble

Acknowledgments

We acknowledge support of the Department of Atomic Energy, Government of India, under Project Identification No. RTI 4007. We sincerely acknowledge Tata Institute of Fundamental Research Hyderabad, India for providing the support of computing resources. We thank to D. E. Shaw Research for providing us the long MD simulation trajectories of Trp-cage, BPTI, Ash1, and monomeric α-Synuclein.[10,68] JM acknowledges Core Research grants provided by the Department of Science and Technology (DST) of India (CRG/2023/001426).

References

(1) Wright, P. E.; Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology* **1999**, *293*, 321–331.

(2) Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS letters* **2005**, *579*, 3346–3354.

(3) Holehouse, A. S.; Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology* **2024**, *25*, 187–211.

(4) Menon, S.; Mondal, J. Conformational plasticity in α-synuclein and how crowded environment modulates it. *The Journal of Physical Chemistry B* **2023**, *127*, 4032–4049.

(5) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* **2008**, *37*, 215–246.

(6) Breydo, L.; Wu, J. W.; Uversky, V. N. α-Synuclein misfolding and Parkinson's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **2012**, *1822*, 261–285.

(7) Ross, C. A. Polyglutamine pathogenesis: emergence of unifying mechanisms for Huntington's disease and related disorders. *Neuron* **2002**, *35*, 819–822.

(8) Uversky, V. N. Introduction to intrinsically disordered proteins (IDPs). 2014.

(9) Spillantini, M. G.; Schmidt, M. L.; Lee, V. M.-Y.; Trojanowski, J. Q.; Jakes, R.; Goedert, M. α-Synuclein in Lewy bodies. *Nature* **1997**, *388*, 839–840.

(10) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J., et al. Millisecond-scale molecular dynamics simulations on Anton. Proceedings of the conference on high performance computing networking, storage and analysis. 2009; pp 1–11.

(11) Shaw, D. E.; Grossman, J.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H., et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2014; pp 41–53.

(12) Kutzner, C.; Pll, S.; Fechner, M.; Esztermann, A.; de Groot, B. L.; Grubmller, H. Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *Journal of Computational Chemistry* **2015**, *36*, 1990–2008.

(13) Huber, T.; Torda, A. E.; Van Gunsteren, W. F. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *Journal of computer-aided molecular design* **1994**, *8*, 695–708.

(14) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 826–843.

(15) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of computational physics* **1977**, *23*, 187–199.

(16) Kästner, J. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 932–942.

(17) Grubmüller, H.; Heymann, B.; Tavan, P. Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **1996**, *271*, 997–999.

(18) Isralewitz, B.; Gao, M.; Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Current opinion in structural biology* **2001**, *11*, 224–230.

(19) Maragliano, L.; Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chemical physics letters* **2006**, *426*, 168–175.

(20) Abrams, C. F.; Vanden-Eijnden, E. Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proceedings of the National Academy of Sciences* **2010**, *107*, 4961–4966.

(21) Doerr, S.; De Fabritiis, G. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *Journal of chemical theory and computation* **2014**, *10*, 2064–2069.

(22) Betz, R. M.; Dror, R. O. How effectively can adaptive sampling methods capture spontaneous ligand binding? *Journal of chemical theory and computation* **2019**, *15*, 2053–2063.

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

(24) Rhee, Y. M.; Pande, V. S. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophysical journal* **2003**, *84*, 775–786.

(25) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics* **2004**, *120*, 11919–11929.

(26) Perez, D.; Uberuaga, B. P.; Shim, Y.; Amar, J. G.; Voter, A. F. Accelerated molecular dynamics methods: introduction and recent developments. *Annual Reports in computational chemistry* **2009**, *5*, 79–98.

(27) Zuckerman, D. M.; Chong, L. T. Weighted ensemble simulation: review of methodology, applications, and software. *Annual review of biophysics* **2017**, *46*, 43–57.

(28) Huber, G. A.; Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophysical journal* **1996**, *70*, 97–110.

(29) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature communications* **2018**, *9*, 5.

(30) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of chemical physics* **2018**, *148*.

(31) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics* **2011**, *134*.

(32) Husic, B. E.; Pande, V. S. Markov state models: From an art to a science. *Journal of the American Chemical Society* **2018**, *140*, 2386–2396.

(33) Olsson, S.; Wu, H.; Paul, F.; Clementi, C.; Noé, F. Combining experimental and simulation data of molecular processes via augmented Markov models. *Proceedings of the National Academy of Sciences* **2017**, *114*, 8265–8270.

(34) Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **1989**, *77*, 257–286.

(35) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *The Journal of chemical physics* **2013**, *139*.

(36) Eddy, S. R. What is a hidden Markov model? *Nature biotechnology* **2004**, *22*, 1315–1316.

(37) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **2006**, *313*, 504–507.

(38) Liou, C.-Y.; Cheng, W.-C.; Liou, J.-W.; Liou, D.-R. Autoencoder for words. *Neurocomputing* **2014**, *139*, 84–96.

(39) Bera, P.; Mondal, J. Machine learning unravels inherent structural patterns in Escherichia coli Hi-C matrices and predicts chromosome dynamics. *Nucleic Acids Research* **2024**, *52*, 10836–10849.

(40) Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**,

(41) Kingma, D. P.; Welling, M., et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* **2019**, *12*, 307–392.

(42) Degiacomi, M. T. Coupling molecular dynamics and deep learning to mine protein conformational space. *Structure* **2019**, *27*, 1034–1040.

(43) Jin, Y.; Johannissen, L. O.; Hay, S. Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins: structure, function, and bioinformatics* **2021**, *89*, 915–921.

(44) Gupta, A.; Dey, S.; Hicks, A.; Zhou, H.-X. Artificial intelligence guided conformational mining of intrinsically disordered proteins. *Communications biology* **2022**, *5*, 610.

(45) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems* **2014**, *27*.

(46) Brock, A. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096* **2018**,

(47) Rezende, D.; Mohamed, S. Variational inference with normalizing flows. International conference on machine learning. 2015; pp 1530–1538.

(48) Köhler, J.; Krämer, A.; Noé, F. Smooth normalizing flows. *Advances in Neural Information Processing Systems* **2021**, *34*, 2796–2809.

(49) Bond-Taylor, S.; Leach, A.; Long, Y.; Willcocks, C. G. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *44*, 7327–7347.

(50) Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*, eaaw1147.

(51) Hochreiter, S. Long Short-term Memory. *Neural Computation MIT-Press* **1997**,

(52) Sundermeyer, M.; Schlüter, R.; Ney, H. Lstm neural networks for language modeling. Interspeech. 2012; pp 194–197.

(53) Tsai, S.-T.; Kuo, E.-J.; Tiwary, P. Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nature communications* **2020**, *11*, 5115.

(54) Tsai, S.-T.; Fields, E.; Xu, Y.; Kuo, E.-J.; Tiwary, P. Path sampling of recurrent neural networks by incorporating known physics. *Nature Communications* **2022**, *13*, 7231.

(55) Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**,

(56) Bera, P.; Mondal, J. Predicting the Future States of Physicochemical Systems Using Transformer-Based generative Machine Learning Model. *bioRxiv* **2024**, 2024–05.

(57) Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. International conference on machine learning. 2015; pp 2256–2265.

(58) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **2020**, *33*, 6840–6851.

(59) Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **2021**, *34*, 8780–8794.

(60) 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* **2021**, *596*, 583–589.

(61) 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* **2024**, 1–3.

(62) 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* **2021**, *373*, 871–876.

(63) 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., et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100.

(64) Janson, G.; Valdes-Garcia, G.; Heo, L.; Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nature Communications* **2023**, *14*, 774.

(65) Janson, G.; Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLOS Computational Biology* **2024**, *20*, e1012144.

(66) Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. 2015; pp 234–241.

(67) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(68) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences* **2018**, *115*, E4758–E4766.

(69) Wang, Y.; Herron, L.; Tiwary, P. From data to noise to data for mixing physics across temperatures with generative artificial intelligence. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2203656119.

(70) Ward, M. D.; Zimmerman, M. I.; Meller, A.; Chung, M.; Swamidass, S.; Bowman, G. R. Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets. *Nature communications* **2021**, *12*, 3023.

(71) Herron, L.; Mondal, K.; Schneekloth Jr, J. S.; Tiwary, P. Inferring phase transitions and critical exponents from limited observations with thermodynamic maps. *Proceedings of the National Academy of Sciences* **2024**, *121*, e2321971121.

(72) Robustelli, P.; Ibanez-de Opakua, A.; Campbell-Bezat, C.; Giordanetto, F.; Becker, S.; Zweckstetter, M.; Pan, A. C.; Shaw, D. E. Molecular basis of small-molecule binding to α-synuclein. *Journal of the American Chemical Society* **2022**, *144*, 2501–2510.

(73) Mondal, J.; Ahalawat, N.; Pandit, S.; Kay, L. E.; Vallurupalli, P. Atomic resolution mechanism of ligand binding to a solvent inaccessible cavity in T4 lysozyme. *PLoS computational biology* **2018**, *14*, e1006180.

(74) Ahalawat, N.; Mondal, J. Mapping the substrate recognition pathway in cytochrome P450. *Journal of the American Chemical Society* **2018**, *140*, 17743–17752.

(75) Sahil, M.; Singh, T.; Ghosh, S.; Mondal, J. 3site Multisubstrate-Bound State of Cytochrome P450cam. *Journal of the American Chemical Society* **2023**, *145*, 23488–23502.

(76) Hicks, A.; Escobar, C. A.; Cross, T. A.; Zhou, H.-X. Fuzzy association of an intrinsically disordered protein with acidic membranes. *JACS Au* **2020**, *1*, 66–78.

(77) Mancinelli, C. D.; Marx, D. C.; Gonzalez-Hernandez, A. J.; Huynh, K.; Mancinelli, L.; Arefin, A.; Khelashvilli, G.; Levitz, J.; Eliezer, D. Control of G protein–coupled receptor function via membrane-interacting intrinsically disordered C-terminal domains. *Proceedings of the National Academy of Sciences* **2024**, *121*, e2407744121.

(78) Levine, Z. A.; Larini, L.; LaPointe, N. E.; Feinstein, S. C.; Shea, J.-E. Regulation and aggregation of intrinsically disordered peptides. *Proceedings of the National Academy of Sciences* **2015**, *112*, 2758–2763.

(79) Strodel, B. Energy landscapes of protein aggregation and conformation switching in intrinsically disordered proteins. *Journal of molecular biology* **2021**, *433*, 167182.

(80) Benayad, Z.; von Bulow, S.; Stelzl, L. S.; Hummer, G. Simulation of FUS protein condensates with an adapted coarse-grained model. *Journal of chemical theory and computation* **2020**, *17*, 525–537.