Simultaneous Modeling of Protein Conformation and Dynamics via Autoregression

Yuning Shen [*1] **Lihao Wang** [*1] **Huizhuo Yuan** [1] **Yan Wang** [1 2] **Bangji Yang** [1 3] **Quanquan Gu** [1]

Abstract

Understanding protein dynamics is critical for elucidating their biological functions. The increasing availability of molecular dynamics (MD) data enables the training of deep generative models to efficiently explore the conformational space of proteins. However, existing approaches either fail to explicitly capture the temporal dependencies between conformations or do not support direct generation of time-independent samples. To address these limitations, we introduce CONFROVER, an autoregressive model that simultaneously learns protein conformation and dynamics from MD trajectories, supporting both time-dependent and time-independent sampling. At the core of our model is a modular architecture comprising: (i) an *encoding layer*, adapted from protein folding models, that embeds protein-specific information and conformation at each time frame into a latent space; (ii) a *temporal module*, a sequence model that captures conformational dynamics across frames; and (iii) an SE(3) diffusion model as the *structure decoder*, generating conformations in continuous space. Experiments on ATLAS, a large-scale protein MD dataset of diverse structures, demonstrate the effectiveness of our model in learning conformational dynamics and supporting a wide range of downstream tasks. CONFROVER is the first model to sample both protein conformations and trajectories within a single framework, offering a novel and flexible approach for learning from protein MD data.

[*]Equal contribution [1]ByteDance Seed [2]School of Mathematical Sciences, Tongji University, Shanghai [3]Department of Automation, Tsinghua University, Beijing. Correspondence to: Quanquan Gu <quanquan.gu@bytedance.com>.

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

1 Introduction

Proteins are flexible molecules that can adopt multiple structures, called *conformations*. Their ability to transition between different conformations enables biological processes critical to life. Characterizing the behavior of a protein, including its *(1) dynamic motions, (2) conformational distribution, and (3) transitions between different states*, is crucial for understanding its function and guiding the design of novel proteins (McCammon, 1984; Frauenfelder et al., 1991; Berendsen & Hayward, 2000). Molecular dynamics (MD) simulations are the "gold standard" for studying protein conformational changes (McCammon et al., 1977; Childers & Daggett, 2017; Karplus & Kuriyan, 2005). These simulations use physical models to describe the energy of a protein conformation and the forces acting on its atoms. By simulating atomic motion through classical mechanics and iteratively sampling conformations over time, MD enables researchers to explore different conformations, approximating the conformational distribution at equilibrium, and gain mechanistic insights into protein's behavior. However, MD simulations are both computationally expensive and technically challenging due to long simulation times and the tendency to become trapped in local energy minima.

These challenges have motivated the use of deep generative models to study proteins, leveraging the rich conformational and dynamic information provided by large-scale MD datasets (Liu et al., 2024a; Vander Meersche et al., 2024):

(1) *Generating the dynamic motions of proteins is a direct analog to MD simulation.* Pioneering works modeled this by learning transition probabilities of future conformations from the current state (Klein et al., 2024; Schreiner et al., 2024; Costa et al., 2024). However, MD trajectories are often non-Markovian due to partially observed coordinates (e.g., protein-only atoms) and environmental coupling (e.g., using Langevin thermostats). To mitigate this, Cheng et al. (2024) incorporated higher-order information using multiple context frames, though this requires fix context windows and limits flexibility. Jing et al. (2024b) instead modeled the joint distribution over the entire trajectory, capturing complex dependencies among frames. Due to training on fixed-length trajectories and the non-autoregressive design, their model has limited inference-time flexibility that can-

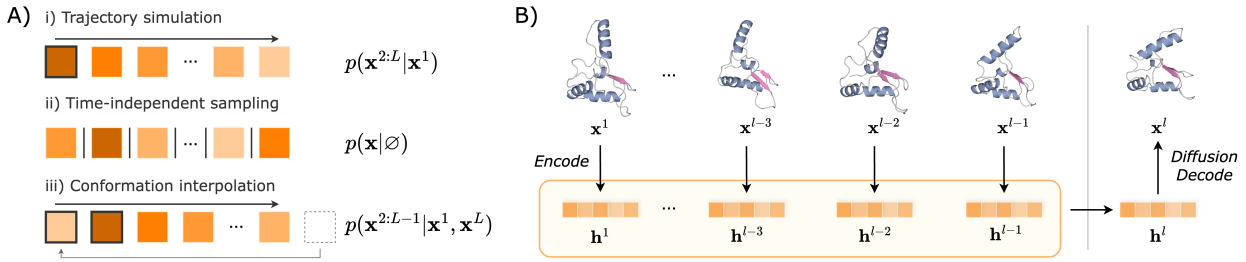

Figure 1: Key ideas of CONFROVER. (A) Various conformation generation tasks are unified as conditional generation, learned via frame-level autoregressive factorization. Each block represents a frame and conditioning frames are outlined in black. The arrow indicates frame dependency. (B) CONFROVER models each frame as a conditional distribution given preceding frames $p(\mathbf{x}^l|\mathbf{x}^{<l})$. Sequential dependencies are captured through latent variables $\mathbf{h}$, and conformations are sampled from a diffusion decoder, conditioned on the updated latent.

not generate variable-length trajectories. Li et al. (2025) introduced an autoregressive approach for flexible trajectory extension, but its deterministic formulation cannot capture trajectory distributions or generate diverse samples.

(2) *Learning the conformational distribution enables sampling time-independent conformations.* Several methods (Noé et al., 2019; Jing et al., 2024a; Wang et al., 2024; Zheng et al., 2024; Lewis et al., 2024) train diffusion- or flow-based generative models on conformation ensembles from MD simulation data, bypassing the need for sequential sampling. While effective for generating samples in parallel, they disregard temporal information in MD trajectories and therefore cannot simulate physical motion of proteins.

(3) *Conformation interpolation generate transition pathways between different states.* Recent works (Du et al., 2024; Jing et al., 2024b) have extended generative modeling to conformation interpolation, where the goal is to generate plausible intermediate samples between known start and end conformational states. Jing et al. (2024b) framed interpolation as a conditional trajectory generation task, but it requires training task-specific model and has not been evaluated on large proteins.

As these generative problems all stem from the same underlying physical principles and involve sampling from a protein's conformational space, a natural question arises: *can we develop a general framework to learn all of these objectives?* We present CONFROVER, a framework for simultaneous learning protein conformation distribution and dynamics from MD trajectory data (Figure 1). Our key observation is that, for an MD trajectory $\mathbf{x}^{1:L}$ of length L, by adopting a general autoregressive formulation, $p(\mathbf{x}^{1:L}) = \prod_{l=1}^{L} p(\mathbf{x}^l|\mathbf{x}^{<l})$, where $\mathbf{x}^{<l}$ denotes all preceding frames of $\mathbf{x}^l$ in the sequence, we can unify multiple generation objectives as instances of frame generation: (1) generating future frame conditioned on all previous frames, suitable for simulating non-Markovian dynamics; (2) unconditional single-frame generation, $p(\mathbf{x}|\varnothing)$, corresponding to time-independent conformation sampling; (3) flexible frame sequence ordering redefines the dependency structure, enabling tasks such as conformation interpolation.

Our contributions are summarized as follows:

- We introduce a simple yet general framework to learn both the conformational distribution and dynamics from MD data, supporting multiple generation tasks including trajectory simulation, time-independent conformation sampling, and conformation interpolation.
- We design a modular architecture that captures temporal dependencies in latent space using efficient causal transformers (i.e., Llama (Touvron et al., 2023)), and directly models conformations in continuous SE(3) space using a diffusion decoder, avoiding discretize structural tokens.
- Experiments show strong capabilities of CONFROVER: it outperforms MDGEN (Jing et al., 2024b) in trajectory simulation, matches the performance of ALPHAFLOW (Jing et al., 2024a) and CONFDIFF (Wang et al., 2024) in time-dependent generation, and can effectively sample conformations interpolating two endpoints.

2 Background

2.1 Data Generation from Molecular Dynamics

Molecular dynamics describes the motion of molecules through Newtonian mechanics $M\ddot{\mathbf{x}} = -\nabla U(\mathbf{x})$, where $\mathbf{x}$ denotes coordinates of atoms in the system, M is the atomic mass, $U(\mathbf{x})$ is the potential energy of the configuration and $-\nabla U(\mathbf{x})$ represents the forces acting on atoms. In practice, stochastic and frictional forces are integrated to model energy exchange with the environment and maintain temperature control of the system, converting the equations of motion to a Langevin process:

$$M\ddot{\mathbf{x}} = -\nabla U(\mathbf{x}) - \gamma M\dot{\mathbf{x}} + \sqrt{2M\gamma k_B T}\boldsymbol{\eta}(t),$$

where γ is the friction coefficient, k_B is the Boltzmann constant, T is the temperature, and $\boldsymbol{\eta}(t)$ is a Gaussian noise term delta-correlated in time $\langle \eta_i(t)\eta_j(t')\rangle = \delta_{ij}\delta(t-t')$. Sampling from this stochastic process generates a time evolution of system configurations. Over time, the samples converges to the Boltzmann distribution $p(\mathbf{x}) \propto$

$\exp(-U(\mathbf{x})/k_BT)$. The trajectory of protein coordinates, $\mathbf{x}_{\text{prot}}^{1:L} = (\mathbf{x}_{\text{prot}}^1, \mathbf{x}_{\text{prot}}^2, \ldots, \mathbf{x}_{\text{prot}}^L)$, is extracted and saved at prescribed simulation intervals, providing both distributional and kinetic information the protein conformational dynamics. For simplicity, we omit the subscript 'prot' and use $\mathbf{x}$ to denote protein coordinates throughout the paper.

2.2 Protein Representations

Proteins are chain-like molecules composed of amino acid *residues*, each selected from 20 standard amino acid types. We parameterize the coordinates of heavy atoms in a protein using the SE(3)-torsional convention (Jumper et al., 2021): the backbone atoms (N−Cα−C) of each residue define a local coordinate via a Gram-Schmidt process, referred to as a *rigid*. The position and orientation of each rigid relative to the global coordinate system are described by a translation-rotation transformation in SE(3) space. The backbone conformation of a protein with N residues can then be represented as $\mathbf{x} = (\mathbf{T}, \mathbf{R}) \in \text{SE}(3)^N$, where $\mathbf{T} \in \mathbb{R}^{N\times 3}$ and $\mathbf{R} \in \text{SO}(3)^N$ are the translation and rotation components. The coordinates of the oxygen atom of the backbone and the side chain atoms can be determined with the addition of up to 7 torsional angles $(\phi, \psi, \omega, \chi_1, \ldots, \chi_4)$ describing the bond rotation. Therefore, the complete configuration of a protein structure is parameterized in the space: $\mathbf{x} = (\mathbf{T}, \mathbf{R}, \phi, \psi, \omega, \chi_1, \ldots, \chi_4) \in (\text{SE}(3) \times \mathbb{T}^7)^N$.

2.3 SE(3)-Diffusion for Protein Conformation Generation

Diffusion generative models are capable of learning complex data distributions. Training involves progressively corrupting data with noise and learning to reverse this process through denoising, thereby modeling the original data distribution (Ho et al., 2020; Song et al., 2021). Recently, diffusion models operating in SE(3) space have been proposed to model protein backbone structures (Yim et al., 2023; Wang et al., 2024). Below, we provide a brief overview of diffusion model and defer the details to Appendix B:

Given a protein backbone conformation as $\mathbf{x}_0 = (\mathbf{T}_0, \mathbf{R}_0) \in \text{SE}(3)^N$, and conditioned on the protein identity (omitted in the equations for clarity), we aim to train a neural network to jointly estimate the score functions of the reverse-time marginal distributions at varying diffusion time t, $s_\theta(\mathbf{x}_t, t) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$. This model is trained using the *denoising score matching* (DSM) loss:

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t, t}\left[\lambda(t)\|s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)\|^2\right].$$

Here $p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ is the forward transition kernel defined in the SE(3) space, $\mathbf{x}_t = (\mathbf{T}_t, \mathbf{R}_t)$ is the noisy data at time t, and $\lambda(t)$ is a time-dependent weight. During inference, DPM generates clean conformations from random noise by simulating the reverse diffusion process with the learned score network $s_\theta(\mathbf{x}_t, t)$.

3 CONFROVER

3.1 Modeling MD Trajectories through Autoregression

Autoregressive generative models factorize the distribution of a sequence as a series of conditional generations over frames. We cast this idea to MD trajectories, modeling a sequence of L frames as:

$$p(\mathbf{x}^{1:L}|\mathcal{P}) = \prod_{l=1}^{L} p(\mathbf{x}^l|\mathbf{x}^{<l}, \mathcal{P}), \tag{1}$$

where $\mathbf{x}^{<l}$ is the preceding frames and $\mathcal{P}$ denotes the protein-specific conditioning input.

Despite its simplicity, this formulation naturally supports multiple learning objectives in protein conformation modeling. In its base form, it models temporal dependencies across the frames, learning to generate the trajectory. When $L = 1$, it removes the frame context and reduces to a single-frame distribution $p(\mathbf{x}|\mathcal{P})$, learning to direct sample time-independent conformations. In addition, the sequential dependency in Equation (1) can be extended to any desired frame-conditioned generation tasks. By prepending conditioning frames $\mathcal{K}$ to the sequence, we train the model to learn any conditional generation $p(\mathbf{x}^{1:L}|\mathcal{K}, \mathcal{P})$, including conformation interpolation by setting $\mathcal{K} = \{\mathbf{x}^1, \mathbf{x}^L\}$. Similar idea was applied in text infilling tasks by shuffling the order of text contexts (Bavarian et al., 2022).

After defining the main learning objectives for trajectory simulation, single-frame (time-independent), and conformation interpolation, we describe how to effectively model autoregressive dependencies over protein conformations in Section 3.2. More critically, in Section 3.3, we explain how to adapt sequence models, traditionally designed for discrete tokens, to the continuous space of protein conformations. Lastly, we introduce a specific choice of architectures of CONFROVER in Section 3.4.

3.2 Latent Causal Modeling

We propose a modular design composed of an encoder, a latent sequence model, and a stochastic decoder. This design enables the use of modern causal transformers, such as in Llama (Touvron et al., 2023), to efficiently capture sequential dependencies between frames in the latent space (Figure 2). During training, the input sequence is shifted by one frame and a mask token "[M]" is prepended; the generation process also begins with the mask token and conditioning frames (e.g., $\mathbf{x}^1$).

To model $p(\mathbf{x}^l|\mathbf{x}^{<l}, \mathcal{P})$, the context frames $\mathbf{x}^{<l}$ are first encoded into intermediate latent states $\mathbf{h}^{<l} = (\mathbf{h}^1, \ldots, \mathbf{h}^{l-1})$ using a shared encoder network with protein-specific condition $\mathcal{P}$:

$$\mathbf{h}^i = f_\eta^{\text{enc}}(\mathbf{x}^i, \mathcal{P}), \quad i = 1, 2, \ldots, l-1. \tag{2}$$

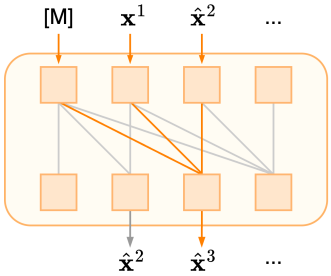

Figure 2: Causal sequence model to generate trajectory $(\hat{\mathbf{x}}^2, \dots)$ from the mask token "[M]" and the conditioning frame $\mathbf{x}^1$. Each frame only attend to its previous frames. Attention activations for $\hat{\mathbf{x}}^3$ are highlighted in orange.

A temporal module is then used to capture the sequential dependencies between the frame latents. Causal attention is used to ensure the latent for frame $\mathbf{h}^l$ is only updated by its preceding frames:

$$\mathbf{h}^l_{\text{updated}} = f^{\text{temp}}_\xi \left(\mathbf{h}^1, \mathbf{h}^2, \dots, \mathbf{h}^{l-1}\right). \tag{3}$$

Since that both the encoder and the temporal module are chosen to be deterministic, $p(\mathbf{x}^l|\mathbf{x}^{<l})$ reduces to a conditional generation over the updated latent, realized through a probabilistic decoder:

$$p(\mathbf{x}^l|\mathbf{x}^{<l}, \mathcal{P}) = p^{\text{dec}}_\theta(\mathbf{x}^l|\mathbf{h}^l_{\text{updated}}). \tag{4}$$

Details on this latent modeling is provided in Appendix C. As a result, we train a model (η, ξ, θ) by jointly optimizing the three modules f^{enc}_η, f^{temp}_ξ, and p^{dec}_θ.

This setup easily accommodates learning single-frame distribution: by replacing all input frames with a mask token and using identity attentions, where each frame only attends to itself, we effectively disable inter-frame information flow. This trains the model to directly sample conformations. A similar strategy has been used in image-video training (Liu et al., 2024b; Ho et al., 2022b).

3.3 Training Autoregressive Model with SE(3) Diffusion Loss

The main challenge in applying autoregressive modeling to conformation trajectories lies in representing the continuous distribution of protein conformations within a framework typically used for discrete token sequences. While some studies have approached this by discretizing the protein structural space into discrete tokens (Hayes et al., 2025; Liu et al., 2023; Lu et al., 2024a), such approaches inherently suffer from discretization error, which can lead to suboptimal performance in modeling protein conformations.

Instead, we propose to directly model the continuous conformational space using diffusion probabilistic models and employ the *DSM loss* for autoregressive model training, similar to Li et al. (2024). Specifically, we perform DSM loss training in SE(3) space: Given a clean frame $\mathbf{x}^l_0 = (\mathbf{T}^l_0, \mathbf{R}^l_0)$, its latent embedding with temporal context $\mathbf{h}^l$ (omit subscript "update" for clarity), the forward transition kernels $p_{t|0}(\mathbf{T}^l_t|\mathbf{T}^l_0)$ and $p_{t|0}(\mathbf{R}^l_t|\mathbf{R}^l_0)$ for the translation and rotation component of SE(3), and a score network to jointly estimate the translation and rotation scores $s_\theta(\mathbf{T}^l_t, \mathbf{h}^l, t)$ and $s^r_\theta(\mathbf{R}^l_t, \mathbf{h}^l, t)$, the loss is defined as

$$\begin{aligned}\mathcal{L}^{\text{SE(3)}}_{\text{DSM}} = &\mathbb{E}\left[\lambda(t)\|s_\theta(\mathbf{T}^l_t, \mathbf{h}^l, t) - \nabla_{\mathbf{T}^l_t} \log p_{t|0}(\mathbf{T}^l_t|\mathbf{T}^l_0)\|^2\right] \\ &+ \mathbb{E}\left[\lambda^r(t)\|s^r_\theta(\mathbf{R}^l_t, \mathbf{h}^l, t) - \nabla_{\mathbf{R}^l_t} \log p_{t|0}(\mathbf{R}^l_t|\mathbf{R}^l_0)\|^2\right],\end{aligned} \tag{5}$$

where the expectation is taken over the diffusion time t and noisy structure $\mathbf{x}^l_t = (\mathbf{T}^l_t, \mathbf{R}^l_t)$ sampled from the forward process. Gradients with respect to $\mathbf{h}^l$ are then backpropagated to update the weights in the temporal module f^{temp}_ξ and encoder f^{enc}_η. During inference, we decode each frame autoregressively by performing reverse sampling as in Equation (6), replacing the scores with estimated values from $s_\theta(\mathbf{T}^l_t, \mathbf{h}^l, t)$ and $s^r_\theta(\mathbf{R}^l_t, \mathbf{h}^l, t)$.

3.4 Model Architecture

An overview of model architecture is shown in Figure 3, with detailed illustrations of each module provided in Appendix D.1.

Encoding Layer. A `FoldingModule`, parameterized by a pretrained OpenFold model (Ahdritz et al., 2022), extracts protein-specific embeddings $\mathcal{P}$ consisting of a single representation ($\mathbf{s}$) and a pair representation ($\mathbf{z}$), shared across frames. For each frame, a `FrameEncoder`, adapted from the template module used in prior works (Jumper et al., 2021; Jing et al., 2024a), encodes pairwise distance of pseudo–C_β atoms via triangular updates and merges this frame pair representations $\mathbf{z}^l_{\text{Frame}}$ with the protein pair representation $\mathbf{z}$. The resulting frame latent embedding, $\mathbf{h}^l = [\mathbf{s}^l, \mathbf{z}^l]$, are invariant to global translation and rotation of the conformations, and are passed to the *Trajectory Module*. Following causal sequence modeling, a masked frame token "[M]" is introduced by zeroing out the pseudo-C_β pairwise distances to remove structural information.

Trajectory Module. The *Trajectory Module* models structural and temporal dependencies across frames, updateing each frame's embedding based on its preceding frames. It consists of interleaved `StructualUpdate` and `TemporalUpdate` layers that operate on the frame-wise latent embeddings. `StructralUpdate` incorporates *Pairformer* layers, a core architecture in protein structure modeling, to update single and pair embeddings through triangular operations (Ahdritz et al., 2022). `TemporalUpdate` employs a Llama-based causal transformer layer for channel-wise self-attention over the sequence of frame embeddings. Each channel in the single and pair embeddings is updated independently. Frame indices are encoded using Rotary Position Embedding (Su

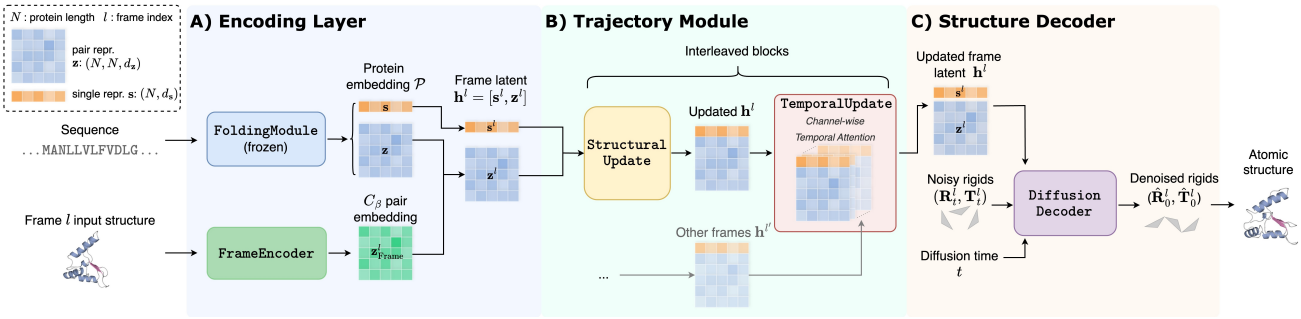

Figure 3: Architecture overview. (A) *Encoding Layer* embeds protein sequence and input structure to each frame as a frame latent representation $\mathbf{h}^l$, comprised of single and pair embeddings; (B) The *Trajectory Module* then updates frame latent $\mathbf{h}^l$ using interleaved structural and temporal update blocks; (C) A diffusion-based *Structure Decoder* learns to denoise a noisy conformation conditioned on the updated frame latent $\mathbf{h}^l$; during inference, it samples conformations from the prior distribution. See Appendix D.1 for details.

et al., 2023). This interleaved design enables efficient updates while maintaining flexibility in modeling sequential dependencies.

Structure Decoder. The updated latent embeddings from *trajectory module* serve as conditioning signals for generating the conformation at each frame. For the SE(3) diffusion model described in Section 2.3, we adopt CONFDIFF (Wang et al., 2024) as the `DiffusionDecoder` to generate 3D conformations. CONFDIFF composes of layers of Invariant Point Attention and Transformer (on single embeddings) to collectively update the residue SE(3) *rigids*, as well as single and pair embeddings. Trained with denoising score matching in Equation (5), the `DiffusionDecoder` learns to iteratively denoise noisy frame structures drawn from a prior SE(3) distribution, conditioned on the frame latent embeddings, to generate accurate backbone conformations of the frame. To reconstruct full-atom geometry, we additionally predict the 7 torsional angles $(\phi, \psi, \omega, \chi_1, \ldots, \chi_4)$ using a light-weight `AngleResNet`, for the coordinates of backbone oxygen atom and side-chain atoms.

4 Experiments

Dataset. We evaluate model performance on ATLAS (Vander Meersche et al., 2024), a large-scale protein MD dataset covering ∼1300 proteins with diverse sizes and structures. For each protein, it contains triplicated 100 ns simulation trajectories. All models are trained on training trajectories and evaluated on test trajectories split by protein identity (Jing et al., 2024a; Wang et al., 2024; Jing et al., 2024b). This presents a challenging task for assessing the generalization to unseen protein structures and dynamics.

Model training. All modules are trained from scratch, except for the `FoldingModule` (frozen OPENFOLD weights) and the `DiffusionDecoder` (initialized from CONFDIFF). During training, sub-trajectories of length $L = 8$ with varying strides (1 ∼ 1024 MD snapshots saved at 10 ps intervals) are sampled to enable learning across multiple timescales. For the main CONFROVER model, we adopt a *hybrid training* strategy using a 1:1 ratio of trajectory and single-frame training objectives; For conformation interpolation task, we continue training the model with a 1:1:1 ratio of trajectory, single-frame and interpolation objectives. See Appendix D.2 for training details.

Baselines. We compare CONFROVER with state-of-the-art deep learning models for each task: For *trajectory simulation*, we compare against MDGEN, a flow-based non-autoregressive trajectory model trained on ATLAS; For *time-independent generation*, we evaluate against ALPHAFLOW and CONFDIFF, flow- and diffusion-based conformation generation models finetuned on ATLAS; For *conformation interpolation*, no existing baseline is available, so we focus on analyzing results of our model.

Remarks on baseline availability. As this is an emerging research area, baseline models remain limited. While recent works on *trajectory simulation* (Klein et al., 2024; Costa et al., 2024; Cheng et al., 2024) and *conformation interpolation* (Du et al., 2024; Jing et al., 2024b) have been proposed, unfortunately, they are not tailored for large proteins as in ATLAS or are not publicly available. See Appendix A for more discussions.

4.1 Trajectory Simulation

Since trajectories from both MD and models are stochastic samples, directly comparing them using frame-wise error, such as root-mean-square-deviation (RMSD) between atomic coordinates, is not appropriate. Therefore, we evaluate the model's ability to recover trajectory dynamics from two perspectives: (1) how well it captures the magnitude of conformation changes across varying start conformations and timescales; (2) how well it recovers the conformational states and principal dynamic modes observed in long-time MD simulations.

Evaluating conformation change accuracy on *multi-start* benchmark. We curated a test benchmark consisting of

Table 1: Pearson correlations of conformation changes between sampled and reference trajectories in *multi-start*. The better scores are highlighted in **bold**.

	Cα coordinates		
	Traj.	Frame	ΔFrame
MDGEN	0.55	0.45	0.40
CONFROVER	**0.77**	**0.63**	**0.53**
	PCA 2D		
	Traj.	Frame	ΔFrame
MDGEN	0.16	0.11	0.10
CONFROVER	**0.75**	**0.50**	**0.43**

short trajectories with $L = 9$ frames, extracted from 82 ATLAS test proteins. For each protein, we choose from varying starting frames (snapshot index $1000 \sim 7000$) and strides ($128 \sim 1024$ snapshots), resulting in a total of $\sim 2,700$ generation conditions. For each trajectory, we measure three aspects of conformational changes: *Trajectory*, the total changes over the entire sequence $\sum_{l=1}^{L-1} d(\mathbf{x}^l, \mathbf{x}^{l+1})$; *Frame*, the changes of each frame relative to the starting frame $d(\mathbf{x}^l, \mathbf{x}^1)$; Δ *Frame*, the changes between consecutive frames $d(\mathbf{x}^l, \mathbf{x}^{l+1})$. Here, $d(\cdot, \cdot)$ measures the distance between two conformations. We report both the L^2-distance in projected 2D PCA space and the RMSD (in Å) of alpha-carbon (Cα) atoms. This benchmark captures diverse dynamics at both the trajectory and frame levels and enables comprehensive evaluation across varying conditions and timescales (see Appendix E.1 for details).

ConfRover shows superior performance in recovering the magnitude of conformational changes. We report the Pearson correlation of measured conformation changes between model-generated and reference trajectories in Table 1, with additional results in Appendix E.1. Compare with MDGEN, CONFROVER shows a significant improvement in correlation scores, mean absolute error and structural quality (Table 7), indicating its stronger ability to recover the magnitude of conformation changes across different starting conditions in the conformational space. The greater difference observed in PCA highlights that CONFROVER more accurately captures conformational changes along the feature dimensions most relevant to the structural variance observed in MD. Figure 4 visualizes ensembles of conformations in generated trajectories, with additional examples in Figure 11. CONFROVER exhibit more notable conformation changes than MDGEN, and reflects the major movements in the structured and loop domains observed in the MD reference.

MDGEN is a non-autoregressive model trained on trajectories of length $L = 250$. Adjusting its inference setup results in degraded conformation. To ensure fair comparison, we use the original inference setting ($S = 40, L = 250$) and

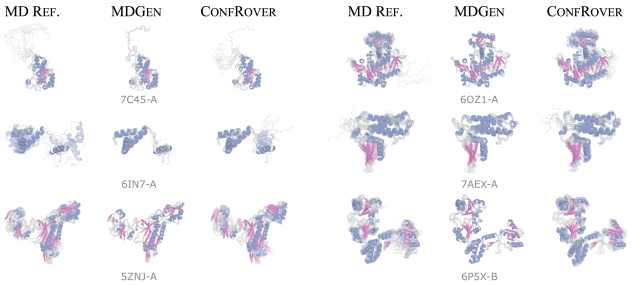

Figure 4: Visualization of six proteins from *multi-start*. Trajectory conformations are colored by their secondary structures and superposed to show the dynamic ensemble. MDGEN primarily exhibits local movements, whereas CONFROVER captures conformations changes similar to MD simulations.

downsampled the trajectories for evaluation. To confirm that this post-processing step does not introduce artifacts, we also trained MDGEN models under the evaluation setups. The results are consistent with the downsampled version (see Appendix E.5).

Table 2: State recovery in 100 ns simulation. Mean values over the first two PCA components are reported. Precision scores for all models are close to 1 and omitted. The better scores are highlighted in bold. MD 100NS is the oracle and excluded in the comparison.

	JSD (↓)	Recall (↑)	F1 (↑)
MD 100NS	0.31	0.67	0.79
MDGEN	0.56	0.30	0.44
CONFROVER	**0.51**	**0.42**	**0.58**

Assessing long trajectory generation on 100 ns simulation. We further evaluate model's ability to recover conformational states and principal dynamics of proteins. For each of protein, we simulate a trajectory of $L = 80$ frames at stride $S = 120$, approximating the 100 ns MD simulation in ATLAS. To assess state recovery, model-generated conformations are projected into a reduced PCA space and compared with the reference trajectory. Specifically, we discretize each principal component into 10 evenly sized "states" and measure the distribution similarity using Jensen-Shannon Distance (JSD). We also compute precision, recall, and F1-score on whether sample conformations fall within these known states (Lu et al., 2024b; Wang et al., 2024; Zheng et al., 2024). To evaluate dynamic mode recovery, we perform time-lagged independent component analysis (tICA) at varying lag times on both reference and sample trajectories. We then compute Pearson correlations between the per-residue contribution to the leading components, based on the tICA coefficients. See Appendix E.2 for details. We include one of the triplicate MD trajectories—excluded from ground-truth evaluation—as an "oracle"

Table 3: Time-independent generation benchmark. ALPHAFLOW and CONFDIFF are specialized SOTA models. The best scores are highlighted in **bold** and the second-best scores are underlined. *MDGEN does not support time-independent sampling; its sequential sampling result are evaluated.

	Pairwise RMSD $r(\uparrow)$	Per target RMSF $r(\uparrow)$	RMWD $(\downarrow)$	MD PCA $\mathcal{W}_2(\downarrow)$	Joint PCA $\mathcal{W}_2(\downarrow)$	Weak contacts J $(\uparrow)$	Transient contacts $J(\uparrow)$	Exposed residue $J(\uparrow)$
ALPHAFLOW	0.53	**0.85**	2.64	1.55	2.29	0.62	**0.41**	**0.69**
CONFDIFF	**0.59**	**0.85**	2.75	1.41	**2.27**	**0.63**	0.39	0.65
MDGEN*	0.41	0.74	2.81	1.95	2.38	0.50	0.28	0.57
CONFROVER-TRAJ	0.47	**0.85**	2.83	1.41	2.30	0.53	0.36	0.60
CONFROVER	0.51	**0.85**	**2.62**	**1.39**	2.28	0.61	0.38	0.67

reference, denoted as MD 100NS, representing the performance expected if the model were as accurate as an MD simulation run.

ConfRover recovers more conformational states than MDGen and accurately captures the principal dynamics. As shown in Table 2, CONFROVER outperforms MDGEN in state recovery, achieving lower JSD, higher recall and F1 scores, showing its improved ability to capture diverse conformations. Additionally, CONFROVER shows clear advantage in capturing the principal dynamic modes across varying lag times, performing even comparably to the MD oracle (Figure 5A). This results suggest CONFROVER can learn and generalize dynamics to unseen proteins and still capture the most important dynamic modes. We visualize simulated trajectories in the PCA space in Figure 5B and Figure 12. These examples again confirm that CONFROVER is more capable of sampling over the conformational space of the proteins and covering diverse conformations. Yet, we also observe some cases where MD 100NS overcame the energy barrier and achieved more remote states while CONFROVER did not (e.g., 7NMQ-A in Figure 5B).

Summary. These experiments demonstrate that CONFROVER outperforms the current state-of-the-art model in trajectory simulation, effectively learning protein dynamics from MD data and generalizing well to unseen proteins. While a gap remains compared to the oracle MD 100NS, particularly in state recovery, the improvement narrows the gap between deep generative models and established simulation methods.

4.2 Time-independent Conformation Sampling

We evaluate the time-independent sampling performance of CONFROVER, following the benchmark setup in Ye et al. (2024). For each protein, 250 independent conformations are generated and compared to MD reference ensembles. Key metrics are summarized in Table 3 with full results in Appendix E.3.

ConfRover matches the performance of state-of-the-art specialized models. Compared to ALPHAFLOW and CONFDIFF, both specialized for time-independent conformation sampling, CONFROVER achieves overall comparable performance and outperforms at least one of the SOTA models in five evaluation criteria. This demonstrate that CONFROVER, despite being a general-purpose model capable of trajectory generation, also performs strongly in sampling independent conformations that approximate the equilibrium distribution from MD simulation. In contrast, MDGEN, which is trained solely for trajectory generation, shows suboptimal results with sequentially sampled conformations.

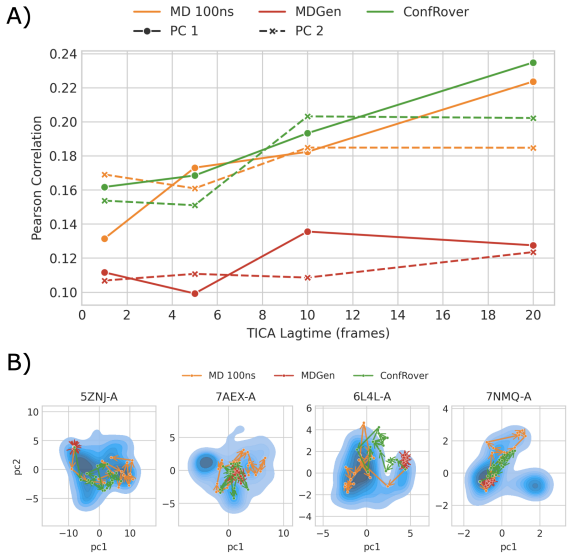

Figure 5: Results from 100 ns simulation. (A) Correlations of principal dynamic modes between sample and reference trajectories, evaluated at varying lag time. (B) Examples trajectories illustrating the states explored by different methods (downsampled by 5 frames for visualization). The blue background indicates the density of the ground-truth conformation distribution from MD reference.

Effect of hybrid training. Without explicit single-frame training, the model primarily learns time-dependent generation, with only the first frame of each trajectory learning to generate conformation unconditionally (i.e., from a masked token input). To test the importance of hybrid training, we ablated the single-frame objective and trained a variant,

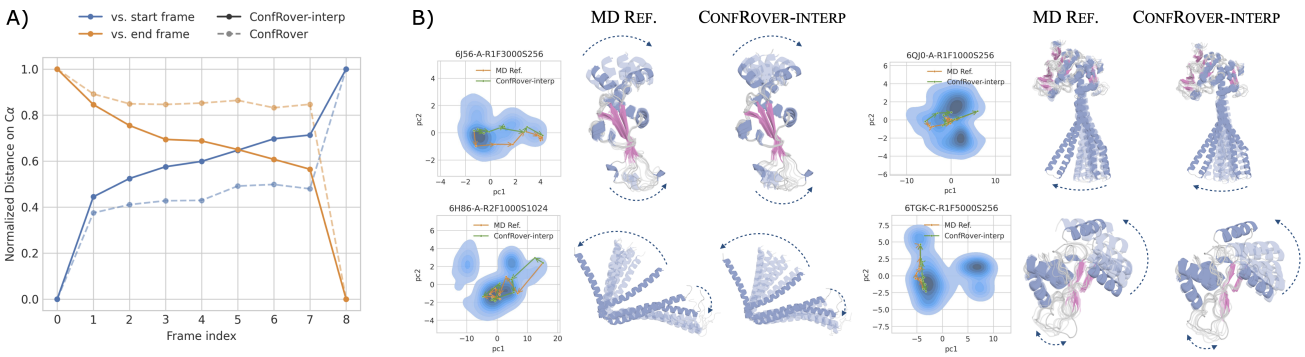

Figure 6: Results from conformation interpolation. (A) Cα-RMSD distance of intermediate frames to the start and end frames, normalized by the distance between start and end frames. Reported values are averaged over 38 cases selected from the *multi-start* benchmark. (B) Example interpolations results. CONFROVER-INTERP generates smooth pathways between the start and end frames, capturing the dynamics observed with the MD reference. Start and end frames are shown as solid structures; intermediate conformations are shown in fading colors. Main motions are indicated by blue dashed arrows.

CONFROVER-TRAJ, solely on trajectory generation. As shown in Table 3, while this variant still outperforms time-dependent results from MDGEN, it shows decreased performance across several metrics compared to CONFROVER. This highlights the importance of hybrid training in balancing the learning objectives and enhancing the model's for generating independent conformations.

4.3 Conformation Interpolation

To enable CONFROVER for conformation interpolation, we continue training CONFROVER with a hybrid objective combining trajectory, single-frame and interpolation, referred as CONFROVER-INTERP. We select 38 short trajectories from *multi-start* for this evaluation where the reference MD trajectories exhibit clear conformation changes and state transitions, see Appendix E.4 for details. To condition on both start and end frames, we prepend the end frame to the start frame and autoregressively generate the remaining (intermediate) frames. To evaluate whether the model generate smooth transitions towards the target end state, we measure Cα-RMSD and L^2-distance in the PCA space between each intermediate frame and the start/end frames.

Training on the interpolation objective enables smooth interpolation between conformations. As shown in Figure 6A, the distance to the start frame increases while the distance to the end frame decreases with frame index, indicating smooth and directed transitions. Without explicit interpolation training, the original CONFROVER (dashed lines in Figure 6A) generates trajectory that do not progress towards the end state. Figure 6B visualizes intermediate structures and transition pathways in PCA space, showing that intermediate conformations from CONFROVER-INTERP closely resemble those in the MD reference. In contrast, as shown in Figure 14, the original CONFROVER can miss key transitions and fails to reach the end state. Additional results and visualizations are provided in Appendix E.4. These results highlight the effectiveness of our interpolation training strategy: by adjusting the dependency order in the sequence model, CONFROVER-INTERP learns to generate smooth transitions between two conformations.

5 Conclusions and Limitations

We introduce CONFROVER, a general framework for learning protein conformational dynamics from MD trajectory data. Through autoregressive factorization, CONFROVER supports three tasks in a unified manner: trajectory simulation, time-independent sampling, and conformation interpolation. This formulation reflects the temporal nature of MD while naturally encompassing conditional and unconditional frame-level generation. Extensive experiments and analyses highlight several empirical advantages: (1) CONFROVER outperforms the current state-of-the-art in trajectory simulation, accurately capturing the magnitude of dynamics, conformational state recovery, and principal motions; (2) Despite being a multi-purpose model, it achieves competitive performance in time-independent sampling compared to specialized methods; (3) With simple sequence reordering training, CONFROVER learns to generate intermediate conformations to interpolate between two given states.

Nevertheless, several limitations remain: (1) Trajectory simulation and interpolation are emerging tasks with limited baselines. We hope this work encourages further development and open-sourcing of models for comprehensive benchmarking; (2) Some evaluation metrics are still preliminary. While interpolation results are intuitive, they would benefit from more meaningful benchmarks reflecting functional state transitions. Evaluation is also constrained by the ATLAS dataset, which is limited to 100 ns simulations and mainly local dynamics. (3) Although CONFROVER narrows the gap with classical MD, it still falls short in fully capturing the conformational space. Future gains may come from scaling training data, using more efficient archi-

tectures, and leveraging additional information from MD, such as energy information. (4) Finally, while the triangular updates in the structural modules ensure high accuracy, their computational cost limits scalability to larger proteins and longer trajectories. Despite these challenges, CONFROVER demonstrates the promise of autoregressive models in molecular simulation, offering a unified, efficient, and extensible approach to modeling protein dynamics.

6 Impact Statement

This work aims to advance machine learning for protein modeling, with broad applications in biology and drug discovery. By enabling efficient simulation of protein dynamics and conformational changes, our method has the potential to accelerate research in structural biology and therapeutic development. Beyond proteins, the approach may also be adapted to other domains of computer-aided design, including small molecule design, materials science, and chip design. While the potential benefits are significant, it is important to ensure responsible use and prevent misuse of the technology. Developing appropriate safeguards and regulatory frameworks will be essential to mitigate any potential negative impacts.

References

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., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493—-500, 2024. doi: 10.1038/s41586-024-07487-w.

Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M. M., Zhang, S., Ojewole, A., Guney, M. E., Biderman, S., Watkins, A. M., Ra, S., Lorenzo, P. R., Nivon, L., Weitzner, B., Ban, Y.-E. A., Sorger, P. K., Mostaque, E., Zhang, Z., Bonneau, R., and AlQuraishi, M. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv*, 2022. doi: 10.1101/2022.11.20.517210. URL https://www.biorxiv.org/content/10.1101/2022.11.20.517210.

Bavarian, M., Jun, H., Tezak, N., Schulman, J., McLeavey, C., Tworek, J., and Chen, M. Efficient training of language models to fill in the middle, 2022. URL https://arxiv.org/abs/2207.14255.

Berendsen, H. J. and Hayward, S. Collective protein dynamics in relation to function. *Current opinion in structural biology*, 10(2):165–169, 2000.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Cheng, K., Liu, C., Su, Q., Wang, J., Zhang, L., Tang, Y., Yao, Y., Zhu, S., and Qi, Y. 4d diffusion for dynamic protein structure prediction with reference guided motion alignment. *arXiv preprint arXiv:2408.12419*, 2024.

Childers, M. C. and Daggett, V. Insights from molecular dynamics simulations for computational protein design. *Mol. Syst. Des. Eng.*, 2:9–33, 2017.

Costa, A. d. S., Mitnikov, I., Pellegrini, F., Daigavane, A., Geiger, M., Cao, Z., Kreis, K., Smidt, T., Kucukbenli, E., and Jacobson, J. Equijump: Protein dynamics simulation via so (3)-equivariant stochastic interpolants. *arXiv preprint arXiv:2410.09667*, 2024.

del Alamo, D., Sala, D., Mchaourab, H. S., and Meiler, J. Sampling alternative conformational states of transporters and receptors with alphafold2. *eLife*, 11:e75751, mar 2022. doi: 10.7554/eLife.75751. URL https://doi.org/10.7554/eLife.75751.

Du, Y., Plainer, M., Brekelmans, R., Duan, C., Noe, F., Gomes, C. P., Aspuru-Guzik, A., and Neklyudov, K. Doob's lagrangian: A sample-efficient variational approach to transition path sampling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ShJWT0n7kX.

Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. The energy landscapes and motions of proteins. *Science*, 254 (5038):1598–1603, 1991.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/abs/10.1126/science.ads0018.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models, 2022b. URL http://arxiv.org/abs/2204.03458.

Jing, B., Erives, E., Pao-Huang, P., Corso, G., Berger, B., and Jaakkola, T. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

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

Jing, B., Stärk, H., Jaakkola, T., and Berger, B. Generative modeling of molecular dynamics trajectories. *arXiv preprint arXiv:2409.17808*, 2024b.

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., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Karplus, M. and Kuriyan, J. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, 102(19):6679–6685, 2005.

Klein, L., Foong, A., Fjelde, T., Mlodozeniec, B., Brockschmidt, M., Nowozin, S., Noé, F., and Tomioka, R. Timewarp: Transferable acceleration of molecular dynamics by learning time-coarsened dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

Köhler, J., Klein, L., and Noé, F. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pp. 5361–5370. PMLR, 2020.

Lewis, S., Hempel, T., Jiménez Luna, J., Gastegger, M., Xie, Y., Foong, A. Y., García Satorras, V., Abdin, O., Veeling, B. S., Zaporozhets, I., et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *bioRxiv*, pp. 2024–12, 2024.

Li, T., Tian, Y., Li, H., Deng, M., and He, K. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

Li, Z., Cen, J., Wu, L., Sun, H., Mao, H., zhangfuzheng, ZHANG, D., and Huang, W. Geometric spatiotemporal transformer to simulate long-term physical dynamics, 2025. URL https://openreview.net/forum?id=LOBhVTtVnc.

Liu, C., Wang, J., Cai, Z., Wang, Y., Kuang, H., Cheng, K., Zhang, L., Su, Q., Tang, Y., Cao, F., Han, L., Zhu, S., and Qi, Y. Dynamic pdb: A new dataset and a se(3) model extension by integrating dynamic behaviors and physical properties in protein structures, 2024a. URL https://arxiv.org/abs/2408.12413.

Liu, H., Liu, S., Zhou, Z., Xu, M., Xie, Y., Han, X., Pérez, J. C., Liu, D., Kahatapitiya, K., Jia, M., et al. Mardini: Masked autoregressive diffusion for video generation at scale. *arXiv preprint arXiv:2410.20280*, 2024b.

Liu, Y., Chen, L., and Liu, H. Diffusion in a quantized vector space generates non-idealized protein structures and predicts conformational distributions. *bioRxiv*, 2023. doi: 10.1101/2023.11.18.567666. URL https://www.biorxiv.org/content/early/2023/11/18/2023.11.18.567666.

Lu, J., Chen, X., Lu, S. Z., Shi, C., Guo, H., Bengio, Y., and Tang, J. Structure language models for protein conformation generation. *arXiv preprint arXiv:2410.18403*, 2024a.

Lu, J., Zhong, B., Zhang, Z., and Tang, J. Str2str: A score-based framework for zero-shot protein conformation sampling. In *The Twelfth International Conference on Learning Representations*, 2024b.

McCammon, J. Protein dynamics. *Reports on Progress in Physics*, 47(1):1, 1984.

McCammon, J. A., Gelin, B. R., and Karplus, M. Dynamics of folded proteins. *nature*, 267(5612):585–590, 1977.

Noé, F., Olsson, S., Köhler, J., and Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Schreiner, M., Winther, O., and Olsson, S. Implicit transfer operator learning: Multiple time-resolution models for molecular dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Stein, R. A. and Mchaourab, H. S. Speach_af: Sampling protein ensembles and conformational heterogeneity with alphafold2. *PLOS Computational Biology*, 18(8):1–16, 08 2022. doi: 10.1371/journal.pcbi.1010483. URL https://doi.org/10.1371/journal.pcbi.1010483.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J.-C., and Galochkina, T. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52(D1):D384–D392, 2024.

Wang, Y., Wang, L., Shen, Y., Wang, Y., Yuan, H., Wu, Y., and Gu, Q. Protein conformation generation via force-guided se (3) diffusion models. In *Forty-first International Conference on Machine Learning*, 2024.

Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., and Kern, D. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, 625(7996):832–839, November 2023. doi: 10.1038/s41586-023-06832-9. URL http://dx.doi.org/10.1038/s41586-023-06832-9.

Ye, F., Zheng, Z., Xue, D., Shen, Y., Wang, L., Ma, Y., Wang, Y., Wang, X., Zhou, X., and Gu, Q. Proteinbench: A holistic evaluation of protein foundation models. *arXiv preprint arXiv:2409.06744*, 2024.

Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. SE(3) diffusion model with application to protein backbone generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 40001–40039, 2023.

Zheng, S., He, J., Liu, C., Shi, Y., Lu, Z., Feng, W., Ju, F., Wang, J., Zhu, J., Min, Y., et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, 6(5):558–567, 2024.

A Extended Discussions

A.1 Related Work

Deep Generative Models for MD Trajectories. Recent works have explored protein trajectory generation as a surrogate for MD simulations. Models such as TIMEWARP (Klein et al., 2024), ITO (Schreiner et al., 2024), and EQUIJUMP (Costa et al., 2024) learn stochastic transport functions to sample future conformations at a lagged time (longer than MD intervals), reducing the computational costs of long-timescale simulations. However, these methods assume Markovian dynamics by relying solely on the current state for prediction, which may not be suitable for non-Markovian dynamics common in protein MD data. To capture higher-order dependencies between the frames, ALPHAFOLDING (Cheng et al., 2024) incorporates history frames via "motion nodes", but it requires a fixed context window. MDGEN (Jing et al., 2024b) takes a different approach by directly modeling the joint distributions of frames in a trajectory and learning frame dependencies through "masked frame modeling", similar to masked language modeling. However, its key-frame parameterization requires separate models for different tasks, and its non-autoregressive paradigm limits flexible generation (e.g., not compatible for generating trajectories with variable lengths). GST (Li et al., 2025) applies autoregression for future frame prediction, enabling variable-length conditioning context and prediction horizons. While the autoregressive approach is conceptually similar to CONFROVER, their work differs in several key aspects: it performs deterministic prediction rather than generative trajectory sampling; it employs a graph-based architecture with fixed structural priors from an adjacency graph, instead of full attention across all residues; it is trained and evaluated on a single protein instead of diverse proteins from ATLAS under a transferable setting.

Beyond forward trajectory simulation, generative models have also been applied to conformation interpolation, that is, generating the intermediate trajectories between two conditioned states. The non-autoregressive framework of MDGEN (Jing et al., 2024b) can be extended to sample transition pathways between such states; however, its key-frame parameterization requires training a separate model for this task, and it has not been tested on large proteins like those in ATLAS. Du et al. (2024) proposed a simulation-free objective for transition-pathway sampling based on Doob's h-transform, but their approach has only been validated on numerical models and the small protein Chignolin. Its generalizability to larger, more diverse proteins remains unassessed.

Notably, the above models focus on learning temporal dependencies between frames and do not support direct, time-independent conformation sampling from the learned distribution. In contrast, CONFROVER is a general framework that learns both the trajectory generation tasks as well as direct sampling of independent protein conformations.

Deep Learning Models for Molecular Conformations. Another line of work focuses on direct sampling of conformations in a time-independent manner. Early efforts include perturbing the input to folding models (e.g., AlphaFold) (del Alamo et al., 2022; Wayment-Steele et al., 2023; Stein & Mchaourab, 2022) or perturbing the conformation using a structural diffusion model (Lu et al., 2024b). However, these models are trained solely on static PDB structures and do not explicitly model the conformational distribution. Recent works have shifted to deep generative paradigms that directly learn protein-specific conformational distributions (Jing et al., 2023; Zheng et al., 2024; Jing et al., 2024a; Wang et al., 2024; Lu et al., 2024a; Lewis et al., 2024). Several models in this category, including ALPHAFLOW, CONFDIFF, and BIOEMULATOR, fine-tune pretrained structure models on large-scale MD datasets, enhancing their ability to capture conformational distributions. A related approach trains normalizing flow models to approximate the Boltzmann distribution (Noé et al., 2019; Köhler et al., 2020), but their invertibility constraints limit scalability and transferability beyond small molecules and peptides. While these methods can generate time-independent conformations, they overlook temporal relationships and do not capture the kinetic aspects of protein dynamics. CONFROVER handles both time-dependent trajectory generation and time-independent conformation learning.

Image-Video Generation. The challenge of modeling protein dynamics conceptually parallels tasks in image and video generation, requiring both data distribution learning and temporal modeling. Recent advances in video generation offer valuable insights into addressing these challenges. Given limited video data, extending image generative models to video has proven effective. Several works (Ho et al., 2022b; Blattmann et al., 2023; Ho et al., 2022a) achieve this by incorporating temporal attention layers, enabling frame-to-frame communication. Disabling temporal attention reverts them to image models, allowing flexible training across both modalities. These approaches efficiently model time correlations without explicitly tracking offsets between frames. Meanwhile, the extension of autoregressive language models to image and video domains has shown strong potential for sequential generation in different data modalities. Li et al. (2024) integrates language models' sequential modeling with diffusion models' ability to model continuous distributions, showing that discrete tokens

are not essential for autoregressive models. MARDINI (Liu et al., 2024b) extends the concept to video generation with efficient llama-style temporal planning and high-resolution video generation via a diffusion decoder. By applying masked "frame" modeling, MARDINI allows the model to learn flexible temporal relationships and enables diverse tasks such as frame interpolation. CONFROVER differs from these works from video models in several aspects: it employs SE(3) diffusion for 3D structure generation; by using an autoregressive paradigm, it explicitly decouples the diffusion generation from temporal modeling, unlike the spatiotemporal denoising process in MARDINI; in addition, the causal autoregressive framework enables more flexible trajectory generation with variable lengths.

A.2 Training and Inference Cost

CONFROVER models in this work contains 19.6 M trainable parameters. All model training and sampling were carried out using 8 NVIDIA H100 GPUs with Distributed Data Parallel. We trained the main CONFROVER model for 180 epochs ($\sim$ 37 hrs) and CONFROVER–INTERP model for additional 220 epochs ($\sim$ 45 hrs).

Inference cost varies with protein size and benchmark scale. For a moderately sized protein (e.g., 300 amino acids), simulating a short trajectory of 9 frames takes less than 1 minute, while generating an 80-frame trajectory (equivalent to 100 ns ATLAS setting in the paper) takes approximately 8.3 minutes. In comparison, a standard MD simulation under the ATLAS setup requires roughly 10 hours to reach 100 ns.

For our large-scale benchmarks, *multi-start* trajectory simulation (2,700+ trajectories) took 8 hours and 30 minutes, while time-independent sampling (250 conformations per protein) took 3 hours and 20 minutes.

A.3 Baseline Limitation

As an emerging research area, few models currently support learning protein dynamics in transferable settings. Existing approaches based on forward transport operators have been primarily trained and evaluated on small peptides (e.g., TIMEWARP (Klein et al., 2024)) or small fast-folding proteins (e.g., EQUIJUMP (Costa et al., 2024)). Although ALPHAFOLDING (Cheng et al., 2024) was trained on ATLAS, its models are not yet publicly available. Due to these limitations, we use MDGEN as the only available model for trajectory-based comparison. For conformation interpolation, neither sampling-based method (Du et al., 2024) nor video-like method MDGEN (Jing et al., 2024b) have been trained and evaluated on large proteins.

B Diffusion Models on SE(3) Space

Diffusion Probabilistic Models (DPM) model complex distributions through iterative denoising. In the context of protein conformations, DPMs defined over $\mathrm{SE}(3)$ translation-rotation space have been applied for protein backbone structural generation (Yim et al., 2023; Wang et al., 2024). Following Section 2.2, $\mathbf{x}_0 = (\mathbf{T}_0, \mathbf{R}_0) \in \mathrm{SE}(3)^N$ denotes the translations and rotations of backbone *rigids* in data. The diffusion processes defined in the translation and rotation subspace add noise to corrupt the data:

$$\begin{aligned}
\mathrm{d}\mathbf{T}_t &= -\frac{1}{2}\beta_t \mathrm{P}\mathbf{T}_t \,\mathrm{d}t + \sqrt{\beta_t}\mathrm{P}\mathrm{d}\mathbf{w}_t, \\
\mathrm{d}\mathbf{R}_t &= \sqrt{\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2}\,\mathrm{d}\mathbf{w}_t^{\mathrm{SO}(3)},
\end{aligned}$$

where $t \in [0, 1]$ is the diffusion time, β_t and σ_t are predefined time-dependent noise schedules and P is a projection operator removing the center of mass. $\mathbf{w}_t$ and $\mathbf{w}_t^{\mathrm{SO}(3)}$ are the standard Wiener processes in $\mathcal{N}(0, I_3)^{\otimes N}$ and $\mathcal{U}(\mathrm{SO}(3))^{\otimes N}$ respectively.

The transition kernel of $\mathbf{T}$ satisfies $p_{t|0}(\mathbf{T}_t|\mathbf{T}_0) = \mathcal{N}(\mathbf{T}_t; \sqrt{\alpha_t}\mathbf{T}_0, (1-\alpha_t)\mathbf{I})$, where $\alpha_t = e^{-\int_0^t \beta_s ds}$. The transition kernel of $\mathbf{R}$ satisfies $p_{t|0}(\mathbf{R}_t|\mathbf{R}_0) = \mathcal{IGSO}_3(\mathbf{R}_t; \mathbf{R}_0, t)$, where $\mathcal{IGSO}_3$ is the isotropic Gaussian distribution on $\mathrm{SO}(3)$ (Yim et al., 2023).

The associated reverse-time stochastic differential equation (SDE) follows:

$$\begin{aligned}
\mathrm{d}\mathbf{T}_t &= \mathbf{P}\left[-\frac{1}{2}\beta_t \mathbf{T}_t - \beta_t \nabla \log p_t(\mathbf{T}_t)\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{P}\mathrm{d}\bar{\mathbf{w}}_t, \\
\mathrm{d}\mathbf{R}_t &= -\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2 \nabla \log p_t(\mathbf{R}_t)\mathrm{d}t + \sqrt{\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2}\mathrm{d}\bar{\mathbf{w}}_t^{\mathrm{SO}(3)},
\end{aligned} \tag{6}$$

where $\bar{\mathbf{w}}_t$ and $\bar{\mathbf{w}}_t^{\mathrm{SO}(3)}$ denote standard Wiener processes in the reverse time.

The reverse process can be approximated by a neural network through the *denoising score matching* loss for translation and rotation:

$$\begin{aligned}
\mathcal{L}(\theta) =& \mathcal{L}^{\mathbf{T}}(\theta) + \mathcal{L}^{\mathbf{R}}(\theta) \\
=& \mathbb{E}\left[\lambda(t)\|s_\theta(\mathbf{T}_t, t) - \nabla_{\mathbf{T}_t}\log p_{t|0}(\mathbf{T}_t|\mathbf{T}_0)\|^2\right] \\
&+ \mathbb{E}\left[\lambda^r(t)\|s_\theta^r(\mathbf{R}_t, t) - \nabla_{\mathbf{R}_t}\log p_{t|0}(\mathbf{R}_t|\mathbf{R}_0)\|^2\right],
\end{aligned} \tag{7}$$

where $\lambda(t)$ and $\lambda^r(t)$ are time-dependent weights, $s_\theta(\mathbf{T_t}, \mathbf{t})$ and $s_\theta^r(\mathbf{R}_t, t)$ are the score networks commonly parameterized with shared weights. The expectations are taken over diffusion time $t \sim \mathcal{U}[t_{\min}, 1]$, and over noisy and clean data pairs from the forward process $(\mathbf{T}_0, \mathbf{T}_t)$ and $(\mathbf{R}_0, \mathbf{R}_t)$.

C Derivation of Equations in Section 3.2

Here we provide a more rigorous derivation of equations in Section 3.2. For clarity, we omit the conditioning variable $\mathcal{P}$ in the intermediate steps.

As defined in Equation (1), our goal is to model the frame-level conditional distribution $p(\mathbf{x}^l|\mathbf{x}^{<l})$. To achieve this, we encode the previous frames $\{\mathbf{x}^i\}_{i=1}^{l-1}$ into a sequence of latent embeddings $\{\mathbf{h}^i\}_{i=1}^{l-1}$, and model inter-frame dependencies in this latent space.

By applying the Bayes' rule, we can factor the joint distribution over the current conformation $\mathbf{x}^l$ and intermediate latent embeddings $\mathbf{h} = (\mathbf{h}^l, \mathbf{h}^{<l})$ as:

$$p(\mathbf{x}^l, \mathbf{h}^l, \mathbf{h}^{<l} \mid \mathbf{x}^{<l}) = p(\mathbf{x}^l \mid \mathbf{h}^l, \mathbf{h}^{<l}, \mathbf{x}^{<l})\, p(\mathbf{h}^l \mid \mathbf{h}^{<l}, \mathbf{x}^{<l})\, p(\mathbf{h}^{<l} \mid \mathbf{x}^{<l}).$$

Integrating both sides over the latent variables $\mathbf{h}$ yields:

$$p(\mathbf{x}^l \mid \mathbf{x}^{<l}) = \int_{\mathbf{h}} p(\mathbf{x}^l \mid \mathbf{h}^l, \mathbf{h}^{<l}, \mathbf{x}^{<l})\, p(\mathbf{h}^l \mid \mathbf{h}^{<l}, \mathbf{x}^{<l})\, p(\mathbf{h}^{<l} \mid \mathbf{x}^{<l})\, \mathrm{d}\mathbf{h}. \tag{8}$$

In our approach, both $p(\mathbf{h}^l \mid \mathbf{h}^{<l}, \mathbf{x}^{<l})$ and $p(\mathbf{h}^{<l} \mid \mathbf{x}^{<l})$ are modeled using deterministic neural networks: an encoder $f_\eta^{\text{enc}}(\mathbf{x}^i)$ and an autoregressive temporal module $f_\xi^{\text{temp}}(\mathbf{h}^{<l})$, respectively.

These mappings reduce both $p(\mathbf{h}^l \mid \mathbf{h}^{<l}, \mathbf{x}^{<l})$ and $p(\mathbf{h}^{<l} \mid \mathbf{x}^{<l})$ to Dirac delta functions, and the conditional dependencies can be simplified as:

$$\begin{aligned} p(\mathbf{h}^l|\mathbf{h}^{<l}, \mathbf{x}^{<l}) &= p(\mathbf{h}^l|\mathbf{h}^{<l}) \\ p(\mathbf{x}^l|\mathbf{h}^l, \mathbf{h}^{<l}, \mathbf{x}^{<l}) &= p(\mathbf{x}^l|\mathbf{h}^l). \end{aligned}$$

Substituting into Equation (8) gives:

$$p(\mathbf{x}^l|\mathbf{x}^{<l}) = \int_{\mathbf{h}} p(\mathbf{x}^l|\mathbf{h}^l) \cdot p(\mathbf{h}^l|\mathbf{h}^{<l}) \cdot p(\mathbf{h}^{<l}|\mathbf{x}^{<l}) \mathrm{d}\mathbf{h}. \tag{9}$$

Again, due to the deterministic nature of the encoder and temporal module, there is no marginalization involved in Equation (9), yielding:

$$\begin{aligned} & p(\mathbf{x}^l|\mathbf{x}^{<l}) = p(\mathbf{x}^l|\mathbf{h}^l) \\ \text{where} \quad & \mathbf{h}^i = f_\eta^{\text{enc}}(\mathbf{x}^i, \mathcal{P}), \quad i = 1, 2, \ldots, l-1 \\ \text{and} \quad & \mathbf{h}^l = f_\xi^{\text{temp}}\left(\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^{l-1}\right). \end{aligned}$$

Finally, we approximate $p(\mathbf{x}^l|\mathbf{h}^l)$ with a parameterized model $p_\theta^{\text{enc}}(\mathbf{x}^l|\mathbf{h}^l)$.

D Method Details

D.1 Detailed Module Architectures

Encoding Layer. The protein-specific single and pair representations are obtained from the Evoformer stack of a pretrained OpenFold model (with frozen weights), after three recycle iterations. In addition, we encode residue-level sequence information by combining sinusoidal positional embeddings of residue indices with learnable embeddings for the 20 standard amino acid types. These features are concatenated with the single representation from the `FoldingModule`.

To encode the structural information of each conformation frame, we introduced the `FrameEncoder`, a pseudo-beta-carbon (C_β) coordinate encoder similar to the `InputEmbedding` module from AlphaFlow (Jing et al., 2024a) (without diffusion time embedding). Specifically, this module first compute the pairwise distances between residues using C_β coordinates. These distances are then binned, embedded into latent embedding, and further refined through triangular update blocks including triangle attention and multiplication updates (Jumper et al., 2021). See Algorithm 1 for the specifics. The resulting per-frame C_β pair embedding $\mathbf{z}^l_{\text{Frame}}$ is concatenated with the pair representation from `FoldingModule`.

Both single and pair embeddings are projected into the same dimension of d for simplicity, forming the latent embedding $\mathbf{h}^l = [\mathbf{s}^l, \mathbf{z}^l]$ for each frame. See detailed illustration in Figure 7.

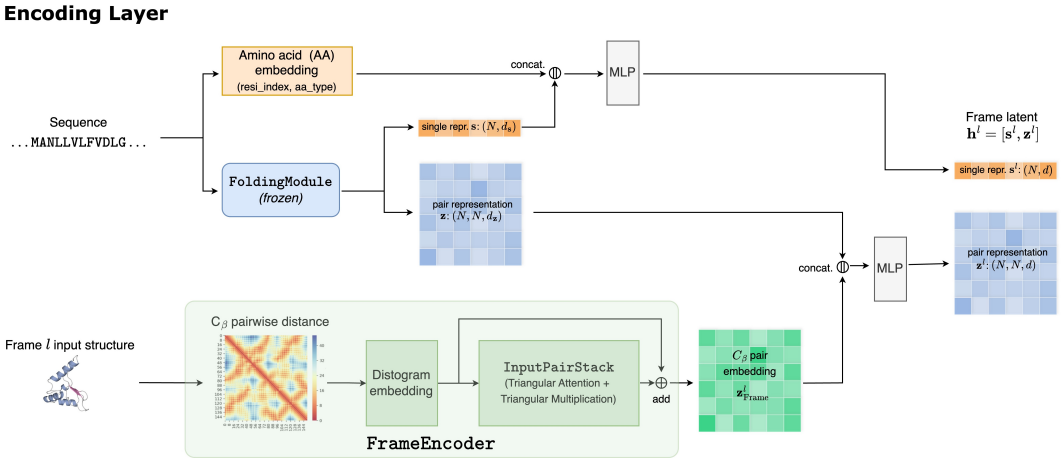

Figure 7: Architecture details of the *Encoding Layer*. A frozen `FoldingModule` encodes the protein-specific information from its sequence, containing prior knowledge on its chemical environment and folding structures. The single representation is further concatenated with additional amino acid embeddings and projected to a hidden dimension of size d; The pair representation is concatenated with frame conformation information, encoded in C_β pair embedding, and projected to a hidden dimension of size d. Both frame-level single and pair embeddings form the frame-level latent for downstream modules.

Algorithm 1 FrameEncoder

Input: Pseudo beta carbon (C_β) coordinates $\mathbf{x} \in \mathbb{R}^{N\times 3}$, time $t \in [0, 1]$
Output: Input pair embedding $\mathbf{z} \in \mathbb{R}^{N\times N\times 64}$
$\mathbf{z}_{ij} \leftarrow \|\mathbf{x}_i - \mathbf{x}_j\|$
$\mathbf{z}_{ij} \leftarrow \text{Bin}(\mathbf{z}_{ij}, \text{min} = 3.25\ \text{Å}, \text{max} = 50.75\ \text{Å}, N_{\text{bins}} = 39)$
$\mathbf{z}_{ij} \leftarrow \text{Linear}(\text{OneHot}(\mathbf{z}_{ij}))$
for $l \leftarrow 1$ to $N_{\text{blocks}} = 4$ **do**
  $\{\mathbf{z}\}_{ij}$ += $\text{TriangleAttentionStartingNode}(\mathbf{z}_{ij}, c = 64, N_{\text{head}} = 4)$
  $\{\mathbf{z}\}_{ij}$ += $\text{TriangleAttentionEndingNode}(\mathbf{z}_{ij}, c = 64, N_{\text{head}} = 4))$
  $\{\mathbf{z}\}_{ij}$ += $\text{TriangleMultiplicationOutgoing}(\mathbf{z}_{ij}, c = 64)$
  $\{\mathbf{z}\}_{ij}$ += $\text{TriangleMultiplicationIncoming}(\mathbf{z}_{ij}, c = 64)$
  $\{\mathbf{z}\}_{ij}$ += $\text{PairTransition}(\mathbf{z}_{ij}, n = 2)$
end for
$\mathbf{z}_{ij} = \text{LayerNorm}(\mathbf{z}_{ij})$

Trajectory Module. In the *Trajectory Module*, we interleave layers of `StructuralUpdate` and `TemporalUpdate` to iteratively update the latent $[\mathbf{s}^l, \mathbf{z}^l]$, enabling the temporal reasoning across frames and structural refinement within each frame.

For the `StructuralUpdate`, we adopt a `Pairformer` block from AlphaFold 3 (Ahdritz et al., 2022), which jointly updates the single and pair embeddings of the current frame through structural reasoning. After the `StructuralUpdate`, the pair embedding is flattened from $[N, N, d]$ to $[N \times N, d]$ and concatenated with the single embedding before being passed into the `TemporalUpdate`.

To model temporal dependencies between frames, we use a lightweight Llama architecture (Touvron et al., 2023). We transpose the input such that the temporal dimension is treated as the sequence axis for channel-wise self-attention across time. Rotary positional encoding (Su et al., 2023) is applied to encode the temporal position for each frame. A causal attention mask is applied to restrict each frame to only attend to previous frames. After the temporal update, the latent embedding are reshaped and split back into single and pair embeddings.

Figure 8 and Table 4 provide the detailed module architecture and hyperparameter configurations, respectively. A `StructuralUpdate` block is included for every two `TemporalUpdate` layers.

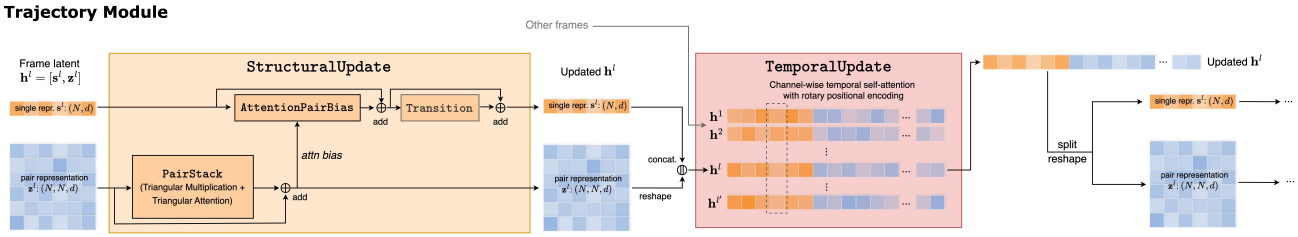

Figure 8: Architecture details of the *Trajectory Module*. *Trajectory Module* contains interleaving blocks of `StructuralUpdate` and `TemporalUpdate` (only one block of each is shown). `StructuralUpdate` leverages the *Pairformer* architecture from Abramson et al. (2024), updating the pair embeddings with triangular updates and the single embeddings using with pair bias from the updated pair. The updated pair embeddings are flatten and concatenate with single embedding for channel-wise temporal update. The attention is applied along the temporal dimension and update each single and pair embedding channels independently. The embeddings from `TemporalUpdate` are split and reshape back into single and pair embeddings.

Table 4: Hyperparameter choices of *Trajectory Module*

Hyperparameters	Values
Lightweight Llama (`TemporalUpdate`)	
Number of layers	8
Dimension of the MLP embeddings	256
Dimension of the hidden embeddings	128
Number of attention heads	4
Pairformer (`StructuralUpdate`)	
Dimension of single embeddings	128
Dimension of pair embeddings	128
Dimension of triangle multiplication hidden embeddings	128
Number of triangle attention heads	4
Dimension of pair attention embeddings	32
Transition layer expanding factor	4
Pair attention dropout rate	0.25

Structure Decoder. Conformation generation, conditioned on the temporal signals from the *Trajectory Module*, is performed using the model architecture from CONFDIFF (Wang et al., 2024), As shown in Figure 9, following CONFDIFF, the inputs to the denoising model include diffusion time t, pairwise distance between residue rigids, and residue indices (not shown). They are encoded and concatenated with the single and pair embeddings from `FoldingModule`. In addition, the single and pair embeddings from the *Trajectory Module* are projected back to the latent dimension $d_{\mathbf{s}}$ and $d_{\mathbf{z}}$, respectively, to modify the single and pair embeddings used by CONFDIFF.

The core of the *Structure Decoder* consists of multiple `IPA-transformer` blocks, which update single and pair embeddings as well as the SE(3) rigids of noisy conformations. In the final block, torsional angles are predicted by `TorsionPred` and, together with denoised rigids, to reconstruct the atomic structure of generated conformation. The corresponding hyperparameter settings are summarized in Table 5.

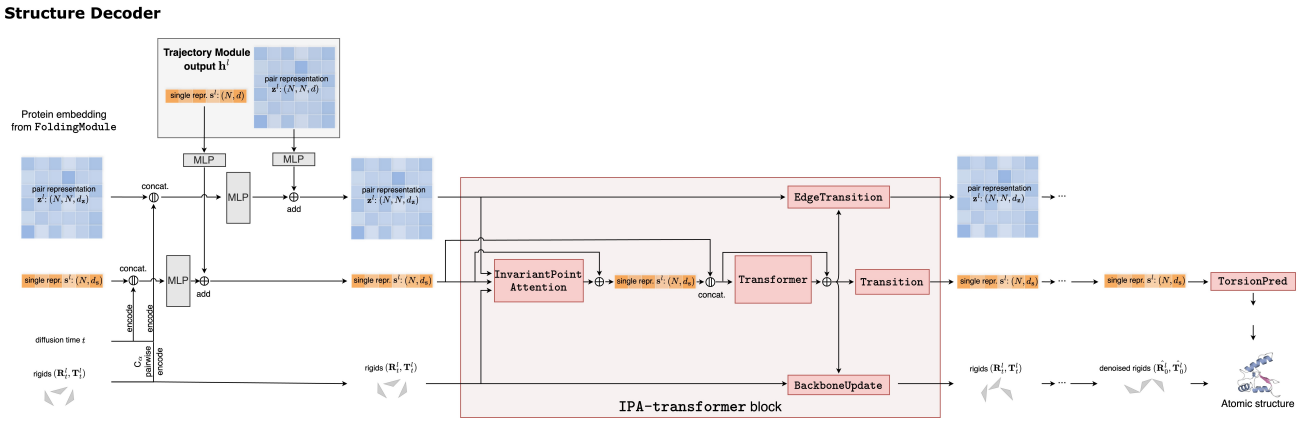

Figure 9: Architecture details of the *Structure Decoder*. Single and pair embeddings from *Trajectory Module* is used to update the original embeddings from `FoldingModule`. The resulting embeddings are fed into blocks of `IPA-transformer` to update single, pair embeddings and denoise rigids, SE(3) representation of protein backbone conformations. Denoised rigids together with torsion angles predicted by `TorsionPred` recovers the atomic structure of protein conformation at this frame.

Table 5: Hyperparameter choices of the *Structure Decoder*

Hyperparameters	Values
Neural network	
Number of IPA blocks	4
Dimension of single embedding ($d_{\mathbf{s}}$)	256
Dimension of pair embedding ($d_{\mathbf{z}}$)	128
Dimension IPA hidden embedding	256
Number of IPA attention heads	4
Number of IPA query points	8
Number of IPA value points	12
Number of transformer attention heads	4
Number of transformer layers	2
SE(3)-diffusion SDE	
Number of diffusion steps	200
Translation scheduler	Linear
Translation $\beta_{\min}$	0.1
Translation $\beta_{\max}$	20
Rotation scheduler	Logarithmic
Rotation $\sigma_{\min}$	0.1
Rotation $\sigma_{\max}$	1.5

D.2 Training and Inference details.

We train all CONFROVER models on the trajectories from the ATLAS training set, following the train-validation-test split of previous works (Jing et al., 2024a; Wang et al., 2024; Jing et al., 2024b). Specifically, we exclude the training proteins longer than 384 amino acid residues, leading to 1080 training proteins.

Most components of CONFROVER models were trained from scratch, except for the `FoldingModule`, where we used frozen weights from OpenFold to extract the single and pair representations from three recycling iterations, and the `DiffusionDecoder`, which was initialized from `ConfDiff-OF-r3-MD` checkpoint provided by the authors[1].

During each training epoch, we randomly sample stride length from 2^0 to 2^{10} to extract sub-trajectories of length $L = 9$ at varying time scales. With the use of causal transformers, input frames were shifted forward by one frame with a `[MASK]` token padded at the beginning of the trajectory. Combined with the use of a causal mask in temporal attention, the design ensures that each frame is trained to sample conditioned only on previous frames and the first frame is generated unconditionally using only the `[MASK]` token as input.

We trained main CONFROVER model 180 epochs ($\sim$ 37 hrs) and CONFROVER–INTERP model for additional 220 epochs ($\sim$ 45 hrs). Additional training hyperparameters can be found in Table 6. All model training and sampling were carried out using 8 NVIDIA H100 GPUs.

Table 6: Training hyperparameters

Hyperparameters	**Values**
Batch Size	1
Frames Num	8
Gradient Clip	1.0
Learning Rate	1×10^{-4}
Optimizer	Adam (weight decay = 0.0)

[1] https://github.com/bytedance/ConfDiff

E Additional Experimental Results

E.1 Trajectory Simulation: *multi-start*

Benchmark Curation. In *multi-start*, we sample short trajectories from varying starting point while ensuring the generation within the scope of the reference trajectory. For example, we select frame index of 1000, 3000, 5000, and 7000 as starting frames for stride $S = 128/256$, resulting in 12 test trajectories from triplicates; frame index 1000, 3000, 5000 as starting frames for stride $S = 512$, and frame index 1000 for stride $S = 1024$ to avoid exceeding total of 10000 frames. This provided us 2,706 different starting conditions from 82 proteins from the ATLAS test set for evaluation.

PCA Projection. Following previous works (Jing et al., 2024a; Wang et al., 2024), we project the Cα coordinates of proteins into a reduced PCA space to focus on the principal dimensions that best capture the structural variations observed in MD simulations. Briefly, for each protein, conformations from triplicate MD simulations in ATLAS are all aligned to the reference conformation (input structure for simulations). The coordinates of each Cα atoms are then flattened and used to fit a per-protein PCA model. For all subsequent analyses, sampled conformation are aligned to the reference structure before computing their PCA projections.

Additional Results. Here we also include the scatterplot of Pearson correlation in Figure 10 and additional metrics from the *multi-start* experiments at different strides in Table 7. Across different strides, CONFROVER models consistently outperforms MDGEN in recovering the correct level of dynamics.

Table 7: Additional metrics from the *multi-start* benchmark. Results for different strides are shown in separate blocks. The better score in each block is highlighted in **bold** (excluding diversity)

	Diversity	MAE on PCA-2D (↓)			MAE on CA coordinates (↓)			Quality
	Pairwise RMSD	Trajectory	Frame	Frame Next	Trajectory	Frame	Frame Next	PepBond Break % (↓)
Stride=128								
MDGEN	1.26	6.53	1.28	0.91	5.02	0.96	0.71	27.9
CONFROVER	1.63	**3.10**	**1.10**	**0.74**	**3.83**	**0.81**	**0.64**	**17.3**
Stride=256								
MDGEN	1.34	7.66	1.54	1.06	6.00	1.14	0.85	27.9
CONFROVER	1.78	**3.91**	**1.28**	**0.87**	**4.60**	**0.91**	**0.75**	**16.6**
Stride=512								
MDGEN	1.40	9.12	1.94	1.25	7.27	1.41	1.01	28.0
CONFROVER	1.89	**4.84**	**1.53**	**1.01**	**5.66**	**1.07**	**0.89**	**16.7**
Stride=1024								
MDGEN	1.51	11.48	2.62	1.55	9.04	1.80	1.24	28.1
CONFROVER	2.04	**6.75**	**1.89**	**1.25**	**7.39**	**1.26**	**1.09**	**16.7**

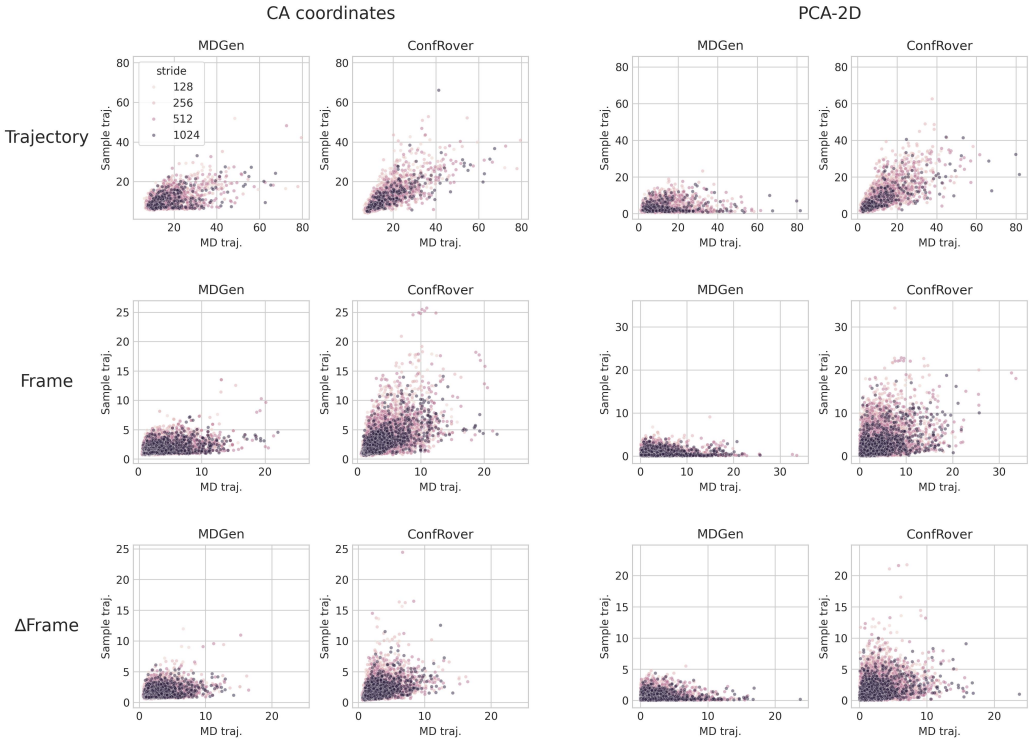

Figure 10: Scatterplots of conformation changes in sample trajectories versus those in the reference trajectories, measured by trajectory-level conformation changes, frame-level conformation changes, and next-frame difference (ΔFrame), measured by the RMSD of alpha carbons (unit: Å) or L^2 distance in the projected PCA space. MDGEN tends underestimate the magnitude of conformation changes while CONFROVER generate samples at similar level as the MD reference. The exact match of measured conformation changes is not possible due to stochastic sampling in both MD simulation and generative models.

Additional Visualization of Trajectory. We additional unfiltered examples (randomly selected) for visual comparison of conformations generated by different models in the *multi-start* experiments, as shown in Figure 11.

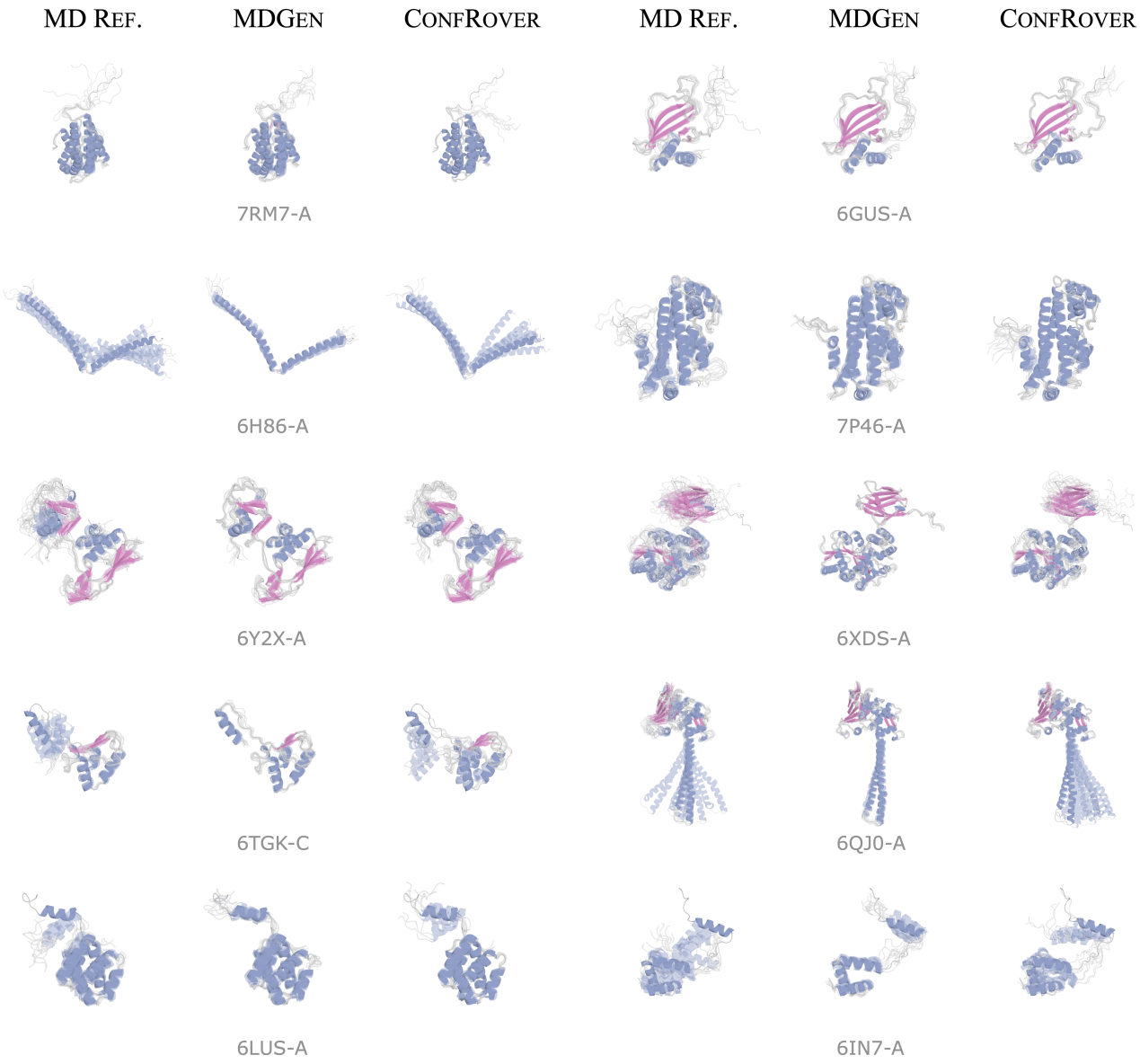

Figure 11: Visualization of 10 trajectories randomly selected from the *Multi-start* benchmark. Trajectory conformations are colored by their secondary structures and superposed to show the dynamic ensemble. MDGEN exhibits primarily local motions while CONFROVER better reflects the motions observed in MD reference.

E.2 Trajectory Simulation: 100 ns Long Trajectory Simulation on ATLAS

Details on Evaluation Metrics. Conformational state recovery is evaluated by comparing the distribution of model-generated and reference conformations in a PCA space. Same as in the *multi-start* benchmark, each conformation is projected into the PCA space parameterized by the 3D coordinates of Cα atoms. To compare distributions, each principal component is discretized into 10 evenly spaced bins. After projecting the conformations into this space, we count their occurrences in each bin and compute the distribution similarity using Jensen–Shannon Distance (JSD). We also binarize the occupancy counts to compute precision, recall, and F1-score—evaluating whether sampled conformations fall within known states, following prior work (Lu et al., 2024b; Wang et al., 2024; Zheng et al., 2024).

Dynamic mode recovery is assessed using time-lagged independent component analysis (tICA) applied separately to reference and generated trajectories across varying lag times. After fitting, we extract tICA coefficients for each Cα atom and compute Pearson correlations between the per-residue contributions to the leading components, evaluating alignment of dynamic modes.

Additional Visualizations. We additionally provide unfiltered (randomly selected) PCA plots in Figure 12.

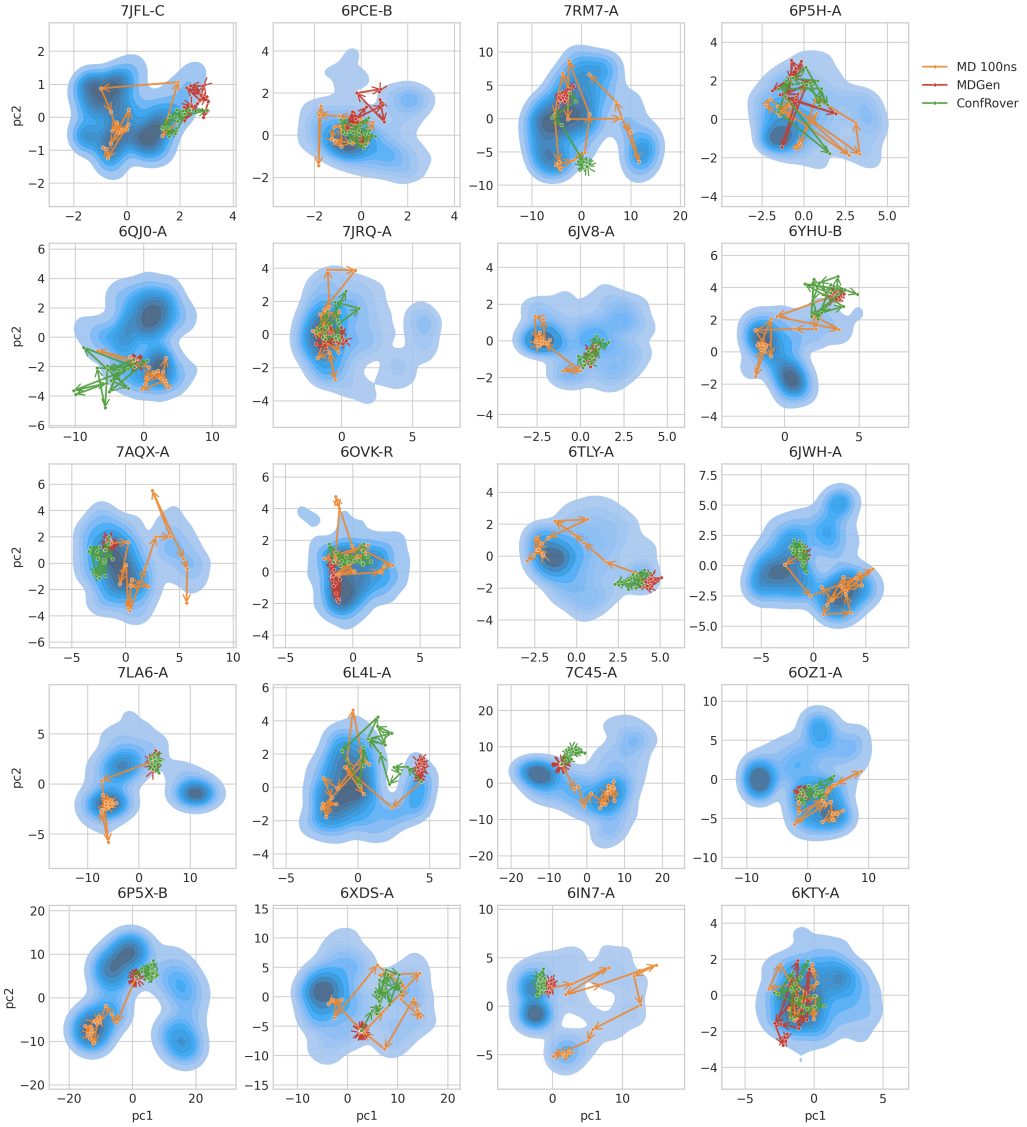

Figure 12: Visualization the ATLAS-100ns trajectories from 20 randomly selected cases. The blue background indicates the density of the ground-truth conformation distribution from MD reference. CONFROVER shows improved conformation state recovery in several cases (e.g., 7JRQ-A, 6YHU-B, 7AQX-A, 6L4L-A, 6OZ1-A, etc), sampling more diverse conformations. Yet, the gap between the oracle MD 100NS and deep learning models is evident in some cases (e.g., 7JFL-C, 6TLY-A, 6JWH-A, etc)

E.3 Time-independent Conformation Sampling

We follow the evaluation protocol of Ye et al. (2024) to assess time-independent conformation sampling on the ATLAS test set. For each protein, 250 independent samples are generated. Since MDGEN does not support time-independent sampling, we approximate its performance using samples from its 100-ns trajectory, serving as a sequential-sampling baseline. The performance of state-of-the-art models, ALPHAFLOW and CONFDIFF, is taken from Table 10 of Ye et al. (2024), using their best-performing variants: ALPHAFLOW-MD and CONFDIFF-OPEN-MD. Full results are shown in Table 8.

Table 8: Performance on time-independent conformation generation on ATLAS. A total of 250 conformations were sampled for each protein, and the median values across 82 proteins are reported. The best performance is highlighted in **bold**, and the second-best is underlined.*MDGEN does not support time-independent sampling and the result from sequential sampling is used.

	Diversity		Flexibility: *Pearson* r on			Distributional accuracy			
	Pairwise RMSD	RMSF	Pairwise RMSD ↑	Global RMSF ↑	Per target RMSF ↑	RMWD ↓	MD PCA $\mathcal{W}_2$ ↓	Joint PCA $\mathcal{W}_2$↓	PC sim > 0.5 %↑
ALPHAFLOW	2.87	1.63	0.53	0.66	**0.85**	2.64	1.55	2.29	39.0
CONFDIFF	3.43	2.21	**0.59**	**0.67**	**0.85**	2.75	1.41	**2.27**	35.4
MDGEN*	1.38	0.78	0.41	0.49	0.74	2.81	1.95	2.38	13.4
CONFROVER-TRAJ	3.21	1.74	0.47	0.61	**0.85**	2.83	1.41	2.30	**41.5**
CONFROVER	3.63	2.23	0.51	0.64	**0.85**	**2.62**	**1.39**	2.28	36.6

	Ensemble observables				Quality	
	Weak contacts J ↑	Transient contacts J↑	Exposed residue J ↑	Exposed MI matrix ρ ↑	CA clash % ↓	PepBond break % ↓
ALPHAFLOW	0.62	**0.41**	**0.69**	**0.35**	**0.0**	22.2
CONFDIFF	**0.63**	0.39	0.65	0.33	0.5	**6.5**
MDGEN*	0.50	0.28	0.57	0.26	0.3	29.0
CONFROVER-TRAJ	0.53	0.36	0.60	0.27	0.4	12.2
CONFROVER	0.61	0.38	0.67	0.32	0.5	19.2

E.4 Conformation Interpolation

In *conformation interpolation* experiment, we selected trajectories from *multi-start*. These trajectories exhibit sufficient conformation changes (e.g., RMSD between the start and end frames > 4 Å) and clear interpolation path in the PCA space. The list of selected cases are in Table 9.

The L^2 distance of generated intermediate frames to the start and end frames in the PCA spaces are reported in Figure 13. Distances are normalized by the distance between start and end frames. Similar to the results measured by Cα-RMSD, CONFROVER-INTERP shows smooth interpolation between the start and end frames while CONFROVER does not. This result shows that by continue training the model on interpolation objective, CONFROVER can learn to generated interpolating trajectories conditioned on the end state.

Table 9: List of 38 selected cases from *multi-init* for interpolation test. Naming conventions: "{PDB_ID}_{Chain_ID}_R{ATLAS repeat}F{Starting index}S{Stride}"

0	5ZNJ-A-R2F1000S512	6L4L-A-R1F3000S256	6TGK-C-R1F5000S256	7JFL-C-R2F3000S256
1	6E7E-A-R3F3000S512	6LRD-A-R1F1000S1024	6TLY-A-R2F1000S256	7JRQ-A-R1F1000S1024
2	6GUS-A-R2F7000S256	6LUS-A-R2F3000S128	6XDS-A-R1F1000S128	7LA6-A-R1F1000S512
3	6H49-A-R2F1000S1024	6OVK-R-R2F7000S128	6XRX-A-R3F5000S128	7LP1-A-R2F5000S256
4	6H86-A-R2F1000S1024	6OZ1-A-R1F1000S512	6Y2X-A-R2F3000S512	7P41-D-R2F3000S256
5	6IN7-A-R3F5000S128	6P5H-A-R1F1000S1024	7AEX-A-R3F5000S256	7P46-A-R3F5000S256
6	6J56-A-R1F3000S256	6P5X-B-R1F3000S256	7AQX-A-R2F3000S512	7RM7-A-R3F3000S128
7	6JPT-A-R2F5000S512	6Q9C-A-R3F3000S256	7ASG-A-R1F7000S128	7S86-A-R3F5000S256
8	6JV8-A-R3F1000S256	6QJ0-A-R1F1000S256	7BWF-B-R3F3000S128	
9	6KTY-A-R3F1000S1024	6RRV-A-R1F1000S256	7C45-A-R1F5000S512	

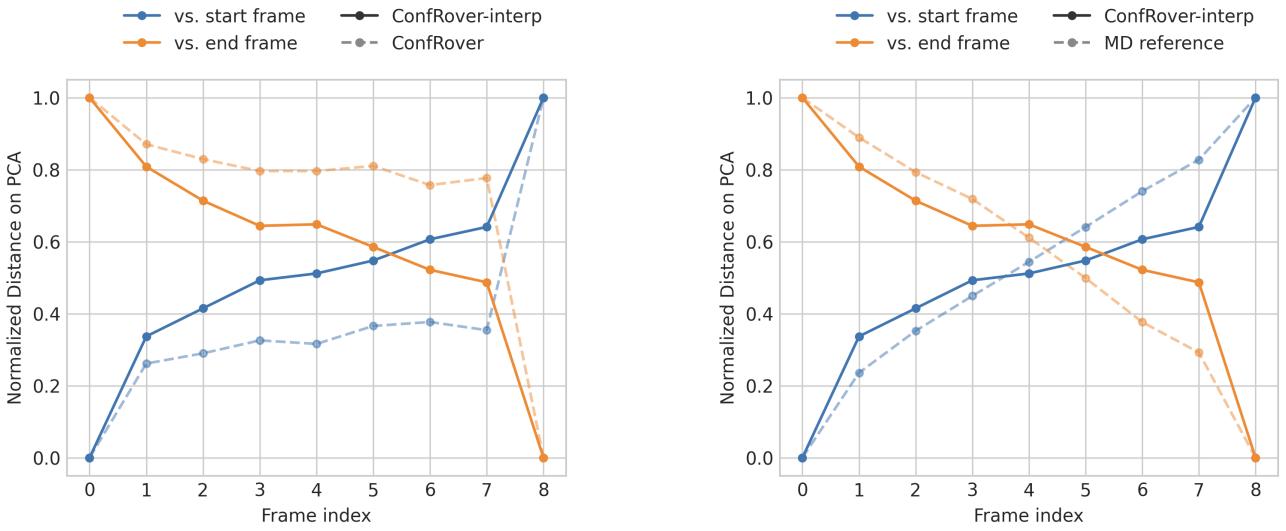

Figure 13: Normalized PCA distance of intermediate frames to the start and end frames, averaged over 38 cases selected from the *multi-start* benchmark. [Left] a comparison between CONFROVER-INTERP and CONFROVER, where CONFROVER-INTERP generates smooth pathways while CONFROVER does not; [Right] a comparison between CONFROVER-INTERP and reference trajectories.

Additional Visualizations. Here we include additional visualization on interpolation results.

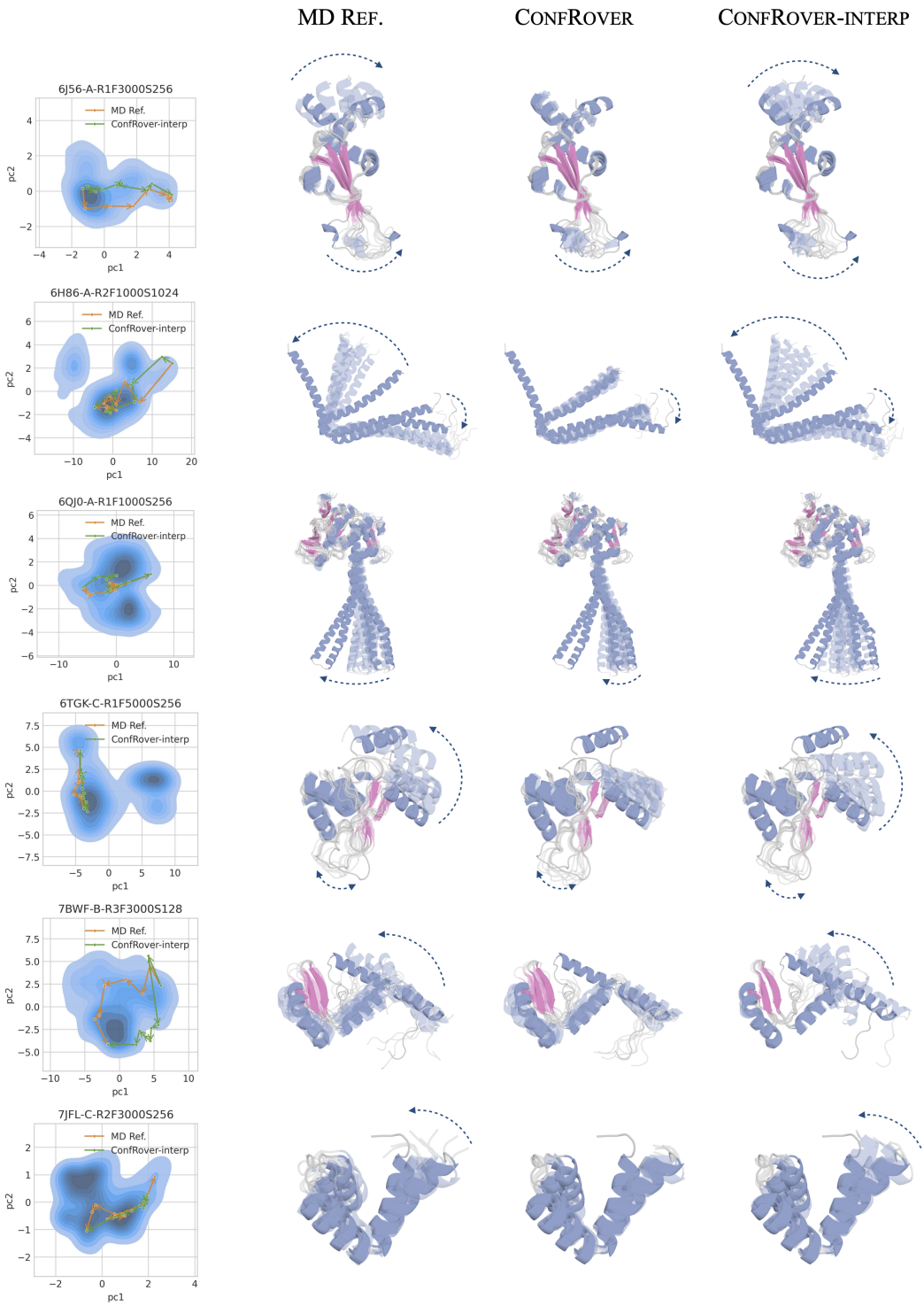

Figure 14: Example interpolations results. CONFROVER-INTERP generates smooth pathways between the start and end frames, capturing the dynamics observed with the MD reference while CONFROVER does not show the correct intermediate conformations. Start and end frames are shown as solid structures; intermediate conformations are shown in fading colors. Main motions are indicated by blue dashed arrows. These examples highlight the difference between the original CONFROVER and CONFROVER-INTERP that further trained on the interpolation objective. The original CONFROVER can miss key motions of the transition while CONFROVER-INTERP correctly capture these motions.

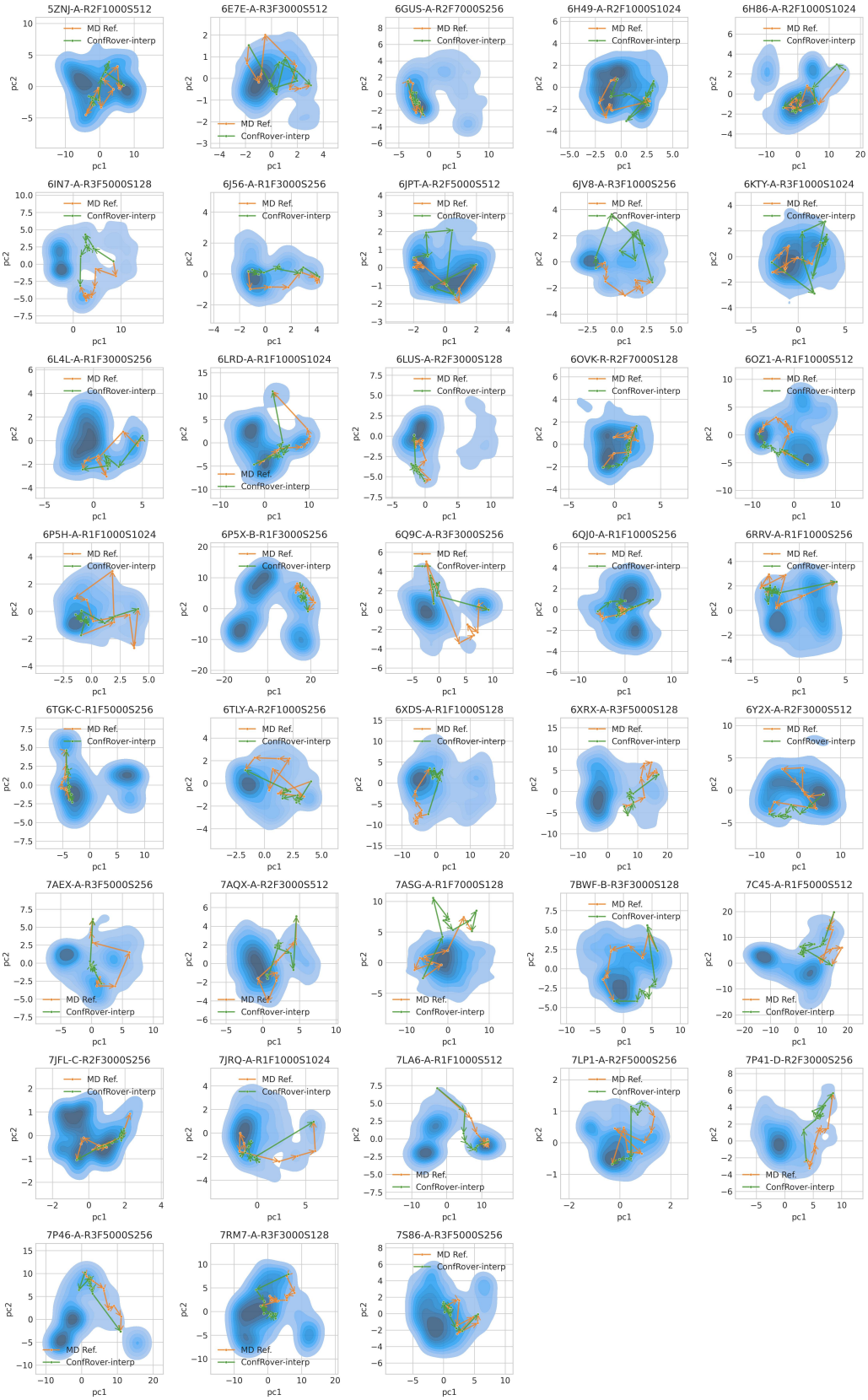

Figure 15: PCA plot of 38 selected interpolation cases. MD reference trajectories and results from CONFROVER-INTERP are shown in each plot.

E.5 Retraining MDGen for Experiment-Specific Setups

Due to its non-autoregressive design and training on fixed-length trajectories, MDGEN cannot directly generate sequences of varying lengths. Therefore, in our evaluation, we generate trajectories using the original settings (named MDGEN-SSTRIDEFLENGTH) and subsample them to match the evaluation setup. However, this post-processing may introduce artifacts. To address this concern, we retrain MDGEN under the evaluation settings and compare the results on the *Multi-start* benchmark (stride = 256) and 100 ns long trajectory simulations, as shown below in Table 10, Table 11, Table 12 and Figure 16. As shown in these experiments, no significant difference of performance for key metrics observed comparing MDGEN with post-processed results and models specifically trained at the evaluation settings, suggesting no evident decrease of trajectory quality from the subsampling post-process.

Table 10: Compare MDGEN-S256F9 with MDGEN from subsampling post-process. Here is the table summarizing the Pearson correlations of conformation changes between sampled and reference trajectories in *multi-start*. MDGEN-S256F9 is trained and sampled with stride of 256 MD snapshots and length of 9 frames. The best scores are highlighted in **bold**.

	Cα coordinates		
	Trajectory	Frame	ΔFrame
MDGEN	0.57	0.46	0.41
MDGEN-S256F9	0.56	0.45	0.38
CONFROVER	**0.77**	**0.62**	**0.53**
	PCA 2D		
	Trajectory	Frame	ΔFrame
MDGEN	0.18	0.13	0.11
MDGEN-S256F9	0.21	0.19	0.11
CONFROVER	**0.75**	**0.5**	**0.44**

Table 11: Compare MDGEN-S256F9 with MDGEN from subsampling post-process. Here is the table summarizing additional metrics in *multi-start* benchmark. MDGEN-S256F9 is trained and sampled with stride of 256 MD snapshots and length of 9 frames. The best scores are highlighted in **bold**.

	Diversity	MAE on PCA-2D (↓)			MAE on CA coordinates (↓)			Quality
	Pairwise RMSD	Traj.	Frame	ΔFrame	Traj.	Frame	ΔFrame	PepBond Break %(↓)
MDGEN	1.34	7.66	1.54	1.06	6.00	1.14	0.85	27.9
MDGEN-S256F9	1.59	6.90	1.47	1.03	5.33	1.13	0.82	**16.2**
CONFROVER	1.78	**3.91**	**1.28**	**0.87**	**4.60**	**0.91**	**0.75**	16.6

Table 12: Compare MDGEN-S256F9 with MDGEN from subsampling post-process. Here is the table summarizing the state recovery performance in 100 ns long trajectory simulation. MDGEN-S120F80 is trained and sampled with stride of 120 MD snapshots and length of 80 frames. The best scores are highlighted in **bold**.

	JSD (↓)	Recall (↑)	F1 (↑)
MD 100NS	0.31	0.67	0.79
MDGEN	0.56	0.30	0.44
MDGEN-S120F80	0.57	0.29	0.42
CONFROVER	**0.51**	**0.42**	**0.58**

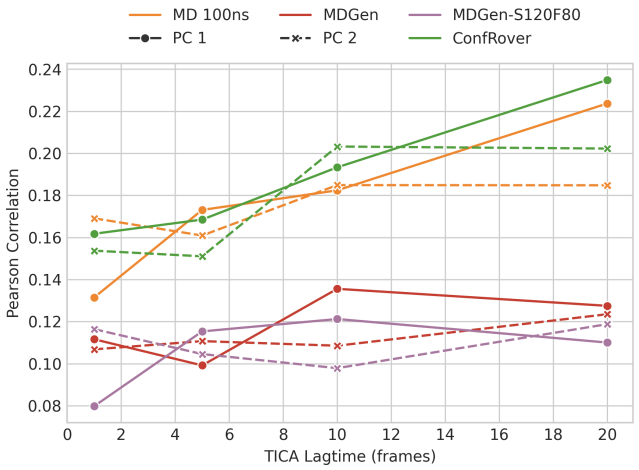

Figure 16: Compare MDGEN-S256F9 with MDGEN from subsampling post-process. This figure shows correlations of main dynamic modes between sampled trajectory and reference trajectory. MDGEN-S120F80 is trained and sampled with stride of 120 MD snapshots and length of 80 frames.

E.6 Extension to Masked Sequence Modeling

In addition to the causal transformer used in CONFROVER, alternative sequence modeling approaches can also be adopted for autoregressive generation. Similar to Li et al. (2024), sequence models trained with bidirectional attention and masked sequence modeling can perform autoregressive generation by iteratively predicting frames in a specified order. At each iteration, the generated frames replace their corresponding masks and are used as input for the next step. We trained a variant CONFROVER-MASK following this approach. Specifically, we replaced causal attention with bidirectional attention in the transformer and implemented a scheduled masking strategy for training, linearly increasing the mask rate from 45% to 88%.

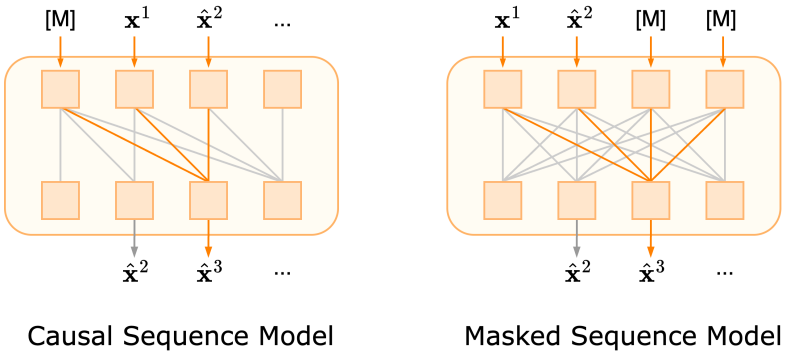

Figure 17: Illustration of autoregressive generation using causal and masked language modeling. "[M]" denotes the mask token and $\mathbf{x}^1$ is the initial conditioning frame for trajectory simulation. Attention activations for the current predicting frame $\hat{\mathbf{x}}^3$ are highlighted in orange. For masked sequence model, the entire sequence is provided as the input with unpredicted frames masked, the frames are generated iteratively to replace the input frames in the future iteration.

As shown in Table13, this masked version performed slightly worse than the CONFROVER (causal) model in recovering the magnitude of conformation changes from various initial conditions, although both clearly outperform MDGEN. These results suggest the potential to extend our framework to broader sequence modeling approaches for autoregressive generation.

However, we also observed several practical limitations with masked sequence modeling. Similar to previous non-autoregressive model, CONFROVER-MASK lacks flexibility to generate trajectories of variable lengths. We found structural degradation when inferring with lengths greatly different from the training setup, such as setting $L = 1$ for time-independent conformation sampling. As a result, it is not suitable for generating long trajectories without resorting to workarounds such as sliding window prediction with window sizes consistent with training window size. While training improvements, such as varying training window sizes, might mitigate these issues, we leave such exploration for future work.

Table 13: Results of CONFROVER-MASK on the Pearson correlations of conformation changes between sampled and reference trajectories in *multi-start*. The better scores are highlighted in **bold**.

	Cα coordinates		
	Traj.	Frame	ΔFrame
MDGEN	0.55	0.45	0.40
CONFROVER	**0.77**	**0.63**	**0.53**
CONFROVER-MASK	0.71	0.60	0.49
	PCA 2D		
	Traj.	Frame	ΔFrame
MDGEN	0.16	0.11	0.10
CONFROVER	**0.75**	**0.50**	**0.43**
CONFROVER-MASKED	0.73	0.49	0.41