Protein Conformation Generation via Force-Guided $\mathrm{SE}(3)$ Diffusion Models

Yan Wang [*1 2] **Lihao Wang** [*1] **Yuning Shen** [1] **Yiqun Wang** [1] **Huizhuo Yuan** [3] **Yue Wu** [3] **Quanquan Gu** [1]

Abstract

The conformational landscape of proteins is crucial to understanding their functionality in complex biological processes. Traditional physics-based computational methods, such as molecular dynamics (MD) simulations, suffer from rare event sampling and long equilibration time problems, hindering their applications in general protein systems. Recently, deep generative modeling techniques, especially diffusion models, have been employed to generate novel protein conformations. However, existing score-based diffusion methods cannot properly incorporate important physical prior knowledge to guide the generation process, causing large deviations in the sampled protein conformations from the equilibrium distribution. In this paper, to overcome these limitations, we propose a force-guided $\mathrm{SE}(3)$ diffusion model, ConfDiff, for protein conformation generation. By incorporating a force-guided network with a mixture of data-based score models, ConfDiff can generate protein conformations with rich diversity while preserving high fidelity. Experiments on a variety of protein conformation prediction tasks, including 12 fast-folding proteins and the Bovine Pancreatic Trypsin Inhibitor (BPTI), demonstrate that our method surpasses the state-of-the-art method.

1 Introduction

Proteins are dynamic macromolecules that play pivotal roles in various biological processes. Their functionality is realized primarily through conformational changes–structural alterations that enable proteins to interact with other molecules. Depicting the protein conformational landscape provides vital insights for (1) identifying potential druggable sites hidden beneath the protein surface, and (2) revealing transition pathways between multiple metastable states. A comprehensive understanding of protein conformations facilitates the elucidation of biological reaction mechanisms, thereby empowering researchers to design targeted inhibitors and therapeutic agents with improved specificity and efficacy.

Traditional physics-based simulation methods such as molecular dynamics (MD) simulations have been extensively studied for protein conformation sampling. With a well-designed empirical force field and numerical integrator, the model propagates the 3D structure of a protein system over time following Newtonian mechanics. MD simulations converge towards the equilibrium distribution (i.e., the Boltzmann distribution) given sufficient time, which facilitates estimation of significant thermodynamic properties, e.g., binding free energy change (Wang et al., 2015). However, to preserve energy conservation and ensure numerical stability, the time step of MD simulations is typically only a few femtoseconds. This poses a challenge as certain biological processes of interest, such as protein folding, span much longer timescales, ranging from microseconds to seconds (Shaw et al., 2021). This results in limited sampling efficiency within conventional MD simulations, further compounded by the rare event sampling problem (Trendelkamp-Schroer & Noé, 2016), impeding the research community to widely adopt MD for high throughput studies.

Building upon the cornerstone of powerful folding models (e.g., AlphaFold (Jumper et al., 2021), RoseTTAFold (Baek et al., 2021), OmegaFold (Wu et al., 2022b), etc.), several attempts have been made to tailor these deep neural networks for protein conformation sampling. By perturbing the model input, such as multiple sequence alignment (MSA) masking (Stein & Mchaourab, 2022) or clustering (Wayment-Steele et al., 2023), the folding model provides a more diverse set of possible folded structures, i.e., alternative conformations. However, this heuristic approach cannot guarantee the predicted structure to be a low energy state of the target sequence. More recently, several works have incorporated diffusion models (Song et al., 2020; Ho et al., 2020) for protein conformation generation (Jing et al., 2023; Lu et al., 2024; Zheng et al., 2023). By pretraining on a large amount

[*]Equal contribution [1]ByteDance Research [2]School of Mathematical Sciences, Tongji University, Shanghai (this work was done during Yan's internship at ByteDance Research) [3]Department of Computer Science, University of California, Los Angeles (this work was done during Huizhuo and Yue's internship at ByteDance Research). Correspondence to: Quanquan Gu <quanquan.gu@bytedance.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

of known protein structures and efficient sampling through a predefined stochastic process, these models have shown promise in exploring diverse protein conformational states. Nevertheless, existing diffusion models fall short in utilizing important physical prior information, such as the MD force field, to guide their diffusion process, hampering their capability to faithfully sample diverse protein conformations complying with the Boltzmann distribution.

To address the aforementioned challenges, we propose a novel force-guided diffusion model, CONFDIFF, aiming to generate high fidelity protein conformations that better adhere to the Boltzmann distribution. Drawing inspiration from the contrastive energy prediction (CEP) technique (Lu et al., 2023), we employ the MD energy prior as a *physics-based preference function*. By introducing an additional force guidance network during the diffusion sampling process, it prioritizes generating conformations with lower potential energy, which effectively enhances sampling quality. Our model is trained on general protein structures from the Protein Data Bank (PDB) (Berman et al., 2000) as well as self-generated conformation samples, without relying on MD simulation data (Zheng et al., 2023). To sum up, the main contributions of this work are highlighted as follows:

- Employing a sequence-conditional model to guide the unconditional model, we use classifier-free guidance on $\mathrm{SE}(3)$ to find a better trade-off between conformation quality and diversity. Compared with DiG (Zheng et al., 2023), our method does not rely on MD data during training; compared with Str2Str (Lu et al., 2024), the guidance intensity coefficient provides a higher degree of freedom for balancing sample diversity and quality.
- We utilize the MD energy function as a *physics-based* reward to guide the generation of protein conformations. In addition, we propose an intermediate force guidance strategy during the diffusion sampling process. To the best of our knowledge, this is the first force-guided network suitable for protein conformation generation, contributing to the alignment of diverse conformation predictions with the equilibrium distribution.
- Experiments on a variety of benchmarks demonstrate that our method outperforms the state-of-the-art approaches. In particular, energy and force guidance effectively guide the model to sample conformations with lower energy, leading to diverse samples more truthful to the underlying Boltzmann distribution.

2 Related Work

Protein Conformation Prediction. Perturbing pretrained folding models, such as AlphaFold (Jumper et al., 2021), to obtain a diverse set of alternative conformations marks the first attempt to use deep neural networks for multi-conformation predictions. Stein & Mchaourab (2022) introduced mutations to the MSA representations to obtain different folded structures from AlphaFold. Similarly, Wayment-Steele et al. (2023) clusters MSA by sequence similarity to enable AlphaFold to discover alternative folding states of known metamorphic proteins. Reducing the MSA depth can also unlock multi-conformation prediction capability of AlphaFold (Del Alamo et al., 2022). In addition, Vani et al. (2022) proposed to use the outputs from AlphaFold as initialization for AI-augmented MD simulations. Noé et al. (2019), Janson et al. (2023) and Mansoor et al. (2023) utilized MD simulation data to generate protein conformation ensembles.

Recently, diffusion models have been employed for protein conformation generation. Zheng et al. (2023) proposed Distributional Graphormer (DiG), which is trained on both protein structures from the PDB and MD simulation data. Unlike our proposed method, DiG incorporates an additional regularization into its loss function to align the learned score with MD force field at small diffusion time, which then extends over the entire pathway by the Fokker-Planck equation. Jing et al. (2023) introduced EIGENFOLD, a harmonic diffusion model with physics-inspired prior to sample protein conformations. The cascading-resolution generative process allows efficient sampling across proteins of varying length. The model achieves remarkable performance in a number of benchmark tasks, yet the advantage of using a harmonic prior over the conventional isotropic Gaussian prior is not evidently clear. Inspired by simulated annealing, Lu et al. (2024) proposed Str2Str, a heating-annealing generative framework using an unconditional score model. By adjusting the duration T_δ of heating (i.e., forward diffusion) process, a certain degree of sample diversity could be achieved. However, during the model's training/inference phase, the absence of sequence or energy information evokes a question of whether the diverse generated outcomes adhere to the Boltzmann distribution.

Diffusion Models for Protein Design. Another line of work focuses on developing diffusion models for protein design (Trippe et al., 2022; Anand & Achim, 2022; Wu et al., 2022a). In particular, Watson et al. (2022) repurposed RoseTTAFold (Baek et al., 2021) to generate novel protein-binder backbones with successful experimental validation. Ingraham et al. (2023) proposed Chroma, which introduces a diffusion process that respects the conformational statistics of polymer ensembles, and can be effectively conditioned on protein semantics or even natural language to generate structures with desired properties. Yim et al. (2023b) proposed framediff, an innovative diffusion process on $\mathrm{SE}(3)$ for equivariant protein backbone generation, and has been recently extended to the flow-matching paradigm (Yim et al., 2023a; Bose et al., 2023). Diffusion models have also been applied to antibody sequence-structure co-design (Martinkus et al., 2023; Luo et al., 2022). The effectiveness of these approaches underscores the potential for further advancements of diffusion modeling in protein studies.

Controllable Generation via Guided Diffusion. Controllable generative modeling is key to aligning diffusion models with human preference in many real-world tasks. Both classifier guidance (Dhariwal & Nichol, 2021; Song et al., 2020) and classifier-free guidance (Ho & Salimans, 2022) have been proposed to guide an unconditional model with a preferred conditional variable, showing remarkable performance in a wide range of applications including text-to-image generation (Saharia et al., 2022; Nichol et al., 2021), video generation (Ho et al., 2022), etc. Recently, (Lu et al., 2023) proposed a novel energy guidance policy using a scalar reward function rather than fixed conditioning variables. By optimizing a contrastive energy prediction (CEP) objective, the model is guaranteed to converge to the exact guidance under enough model capacity and data samples. CEP shows great performance in image synthesis and reinforcement learning tasks. However, it is the gradient of energy function that is utilized as guidance in reverse sampling. This inspires us to propose a force guidance strategy which employs an equivariant network to directly approximate the intermediate force vector.

3 Preliminaries

3.1 Protein Backbone Diffusion on SE(3)

We adopt the protein backbone representation from AlphaFold (Jumper et al., 2021): for a protein with N amino acid residues, its backbone atomic coordinates can be parameterized by a collection of N orientation preserving rigid transformations (i.e., frames) to the local [N,C_α,C,O] backbone atoms in each residue. We collectively denote the positions of all N frames by $\mathbf{x}_0 = [\mathbf{T}_0, \mathbf{R}_0] \in \mathrm{SE}(3)^N$, where $\mathbf{T}_0 \in \mathbb{R}^{3N}$ and $\mathbf{R}_0 \in \mathrm{SO}(3)^N$ denote the corresponding translation and rotation operations, respectively. With an additional backbone torsion angle ψ describing the rotation of the oxygen atom around the C–C_α bond within each residue, we can reconstruct the protein backbone structure from the frame representations.

Following Yim et al. (2023b), diffusion modeling on manifold $\mathrm{SE}(3)^N$ is employed for protein backbone generation. Two independent diffusion processes are defined for the translation and rotation subspaces, respectively:

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

where subscript t denotes the diffusion time variable in $[0, 1]$, β_t and σ_t are predefined time-dependent noise schedules, $\mathbf{P}$ is a projection operator removing the center of mass, and $[\mathbf{w}_t, \mathbf{w}_t^{\mathrm{SO}(3)}]$ represents the standard Wiener process in $[\mathcal{N}(0, I_3)^{\otimes N}, \mathcal{U}(\mathrm{SO}(3))^{\otimes N}]$. The transition kernel of $\mathbf{T}$ satisfies $p_t(\mathbf{T}_t|\mathbf{T}_0) = \mathcal{N}(\mathbf{T}_t; \sqrt{\alpha_t}\mathbf{T}_0, (1-\alpha_t)\mathbf{I})$, where $\alpha_t = e^{-\int_0^t \beta_s ds}$. The rotational transition kernel satisfies $p_t(\mathbf{R}_t|\mathbf{R}_0) = \mathrm{IGSO}_3(\mathbf{R}_t; \mathbf{R}_0, t)$, where IGSO_3 is the isotropic Gaussian distribution on $\mathrm{SO}(3)$ (Leach et al., 2022).

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

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

where $[\bar{\mathbf{w}}_t, \bar{\mathbf{w}}_t^{\mathrm{SO}(3)}]$ denotes another standard Wiener process in reverse time.

3.2 Classifier-free Guidance for Diffusion Sampling

Guided sampling has emerged as a critical strategy in developing diffusion models capable of generating samples complying with human instructions. Consider existing paired data $\mathbf{x}_0 \sim p_0(\mathbf{x}_0|c)$ with a conditioning variable c (subscript denotes diffusion time, where $t = 0$ corresponds to the original data), we typically perceive the conditional probability density at time t as $p_t(\mathbf{x}_t|c)$. Applying Bayes' rule we can obtain $p_t(\mathbf{x}_t|c) = \frac{p_t(\mathbf{x}_t)p_t(c|\mathbf{x}_t)}{p(c)}$. Therefore, one can train a classifier to predict the conditioning probability $p_t(c|\mathbf{x}_t)$ with given noisy data $\mathbf{x}_t$, and use the score of the classifier output as guidance (Dhariwal & Nichol, 2021). Instead of training a separate classifier to estimate $\nabla \log p_t(c|\mathbf{x}_t)$, Ho & Salimans (2022) proposed to utilize an implicit classifier $p_t^\gamma(c|\mathbf{x}_t) \propto \left(\frac{p_t(\mathbf{x}_t|c)p(c)}{p_t(\mathbf{x}_t)}\right)^\gamma$, which then leads to a linear combination of an unconditional score estimator $s_\theta(\mathbf{x}_t)$ and a conditional score estimator $s_\theta(\mathbf{x}_t, c)$ to jointly estimate the target score function:

$$\begin{aligned} &\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|c) \\ =&\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \gamma \nabla_{\mathbf{x}_t} \log p_t(c|\mathbf{x}_t) \\ =&\gamma \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|c) + (1-\gamma)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \\ \approx&\gamma s_\theta(\mathbf{x}_t, c) + (1-\gamma)s_\theta(\mathbf{x}_t), \end{aligned}$$

where γ is a hyperparameter controlling the guidance strength. When $\gamma = 0$, it reduces to an unconditional model, while at $\gamma = 1$, it becomes a pure conditional model. These two models can be simultaneously trained under the same hood, where the unconditional model receives masked conditioning variable c during training.

4 Force-Guided Diffusion for Protein Conformation Generation

In this section, we propose force-guided CONFDIFF, a diffusion model targeting multi-conformation generation for proteins. Employing a sequence-based conditional score network to guide an unconditional score model in Section 4.1, CONFDIFF achieves reasonable conformation diversity while ensuring sample quality. Building upon the energy guidance foundations in Section 4.2, a novel force-guided

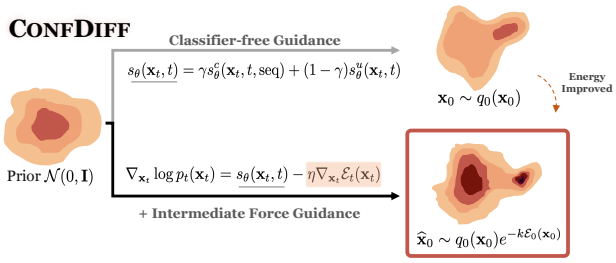

Figure 1. Protein conformation generation with multiple guidance strategies. **Upper:** With a mixture of sequence-conditional and unconditional score models, CONFDIFF in Section 4.1 samples diverse conformations with reasonable quality. **Lower:** Incorporating force guidance in Section 4.3, the model generates structures with lower energy, better comply with the Boltzmann distribution.

sampling strategy is proposed to estimate the intermediate force function, which is then embedded within reverse time sampling process in Section 4.3. Utilizing prior information from the MD force field, our model successfully reweights the generated conformations to ensure they adhere better to the equilibrium distribution. A visual depiction is shown in Figure 1.

4.1 Sequence-Conditional Diffusion on SE(3)

Our baseline model consists of an unconditional score model $s^u_\theta(\mathbf{x}_t, t)$ and a sequence-conditional one $s^c_\theta(\mathbf{x}_t, t, \text{seq})$. The unconditional model is trained on protein structures (from the PDB) without any sequence information, effectively capturing the conformation distribution of general proteins. On the other hand, the sequence-conditional model has access to both protein sequence information (seq) and the corresponding structure $\mathbf{x}_t$ at time t.

We adopt a similar network architecture to FramePred (Yim et al., 2023b) to parameterize the corresponding score functions. The unconditional model takes sinusoidal embedding of the residue index and diffusion time t as its single ($\{\mathbf{s}_i\}$) and pair ($\{\mathbf{z}_{ij}\}$) embeddings, where the conditional model additionally concatenates precomputed representations from ESMFold (Lin et al., 2022) to its single embedding. Note that the choice of sequence representation for the conditional model is flexible – it has been shown that using pretrained representations from folding models helps diffusion models generate reasonable protein structures (Jing et al., 2023; Zheng et al., 2023), while the unconditional model can effectively improve sampling diversity. Both models have been trained with the denoising score matching (DSM) loss function:

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}\left[\lambda(t) \| s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0) \|_2^2\right],$$

where $\lambda(t)$ is a reweighting function inversely proportional to the score norm (Song et al., 2020). A detailed illustration of our model architecture is provided in Appendix B.

During the reverse sampling process, we use a hyperparameter γ to control the classifier-free guidance strength from the conditional model, so that the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \text{seq})$ is estimate by

$$s_\theta(\mathbf{x}_t, t | \text{seq}) = \gamma s^c_\theta(\mathbf{x}_t, t, \text{seq}) + (1-\gamma) s^u_\theta(\mathbf{x}_t, t).$$

For notation simplicity, we hereafter omit the sequence conditional term in our baseline score model, i.e., $s_\theta(\mathbf{x}_t, t) = s_\theta(\mathbf{x}_t, t | \text{seq})$.

4.2 Intermediate Energy-Guided Diffusion

Despite the diffusion model's capability to generate diverse structures, these conformations are not always reasonable in the sense that they may reside in the high energy region of the potential energy surface. Due to the limited availability of multi-conformation data within existing protein structure databases, the training data distribution does not comply with the equilibrium distribution, but rather only containing a few data points near the potential energy minima for each protein sequence. This necessitates the development of a generative model propelled by data and steered by physics-based guidance towards generating samples according to the Boltzmann distribution. To that end, we introduce energy guidance, where the MD energy function attributes a reward to guide the conformation generation process.

Given an existing (baseline) diffusion model which generates samples $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$, our goal is to sample protein conformations from the equilibrium distribution

$$p_0(\mathbf{x}_0) = q_0(\mathbf{x}_0) \frac{e^{-k\mathcal{E}_0(\mathbf{x}_0)}}{Z}, \tag{3}$$

where $Z := \int q_0(\mathbf{x}_0) e^{-k\mathcal{E}_0(\mathbf{x}_0)} \, \mathrm{d}\mathbf{x}_0$ is the intractable normalizing constant, $\mathcal{E}_0$ is the `OpenMM` (Eastman et al., 2017) energy function which evaluates the potential energy of each generated conformation $\mathbf{x}_0$. k is the inverse temperature factor. Given any test function $F(\mathbf{x}_0)$, $p_0(\mathbf{x}_0)$ gives a more accurate estimate through importance sampling, defined as $\mathbb{E}_{p_0(\mathbf{x}_0)}[\frac{F(\mathbf{x}_0)}{e^{-k\mathcal{E}_0(\mathbf{x}_0)}/Z}]$.

In the forward diffusion process, we perturb the original samples $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$ according to the SDE described in Eq. (1), and we assert that $p_t(\mathbf{x}_t | \mathbf{x}_0) := q_t(\mathbf{x}_t | \mathbf{x}_0)$. Whereas, it is the marginal distribution $p_t(\mathbf{x}_t)$ that truly captures our attention. According to Lu et al. (2023), we have the following properties for the energy function:

Proposition 1. *Suppose* $p_0(\mathbf{x}_0) = q_0(\mathbf{x}_0) \frac{e^{-k\mathcal{E}_0(\mathbf{x}_0)}}{Z}$, *and for* $t \in (0, 1]$, $p_t(\mathbf{x}_t | \mathbf{x}_0) := q_t(\mathbf{x}_t | \mathbf{x}_0)$. *Then, the marginal distribution satisfies* $p_t(\mathbf{x}_t) \propto q_t(\mathbf{x}_t) e^{-k\mathcal{E}_t(\mathbf{x}_t)}$, *where* $q_t(\mathbf{x}_t)$ *is the data-based marginal distribution, and* $\mathcal{E}_t(\mathbf{x}_t)$ *satisfies*

$$\mathcal{E}_t(\mathbf{x}_t) = -\frac{1}{k} \log \mathbb{E}_{q_t(\mathbf{x}_0 | \mathbf{x}_t)} \left[e^{-k\mathcal{E}_0(\mathbf{x}_0)}\right]. \tag{4}$$

We refer to $\mathcal{E}_t$ as the intermediate energy function, capturing the dynamic changes in energy at various stages during the diffusion process.

With the exact formula in Eq. (4), a neural network $f_\phi(\mathbf{x}_t, t)$ parameterized by ϕ is utilized to approximate the intermediate energy function $\mathcal{E}_t(\mathbf{x}_t)$ by the optimization problem in Lu et al. (2023)

$$\mathcal{L}_{\text{CEP}} = \mathbb{E}_{p(t)}\mathbb{E}_{q_0(\mathbf{x}_0), q_t(\mathbf{x}_t|\mathbf{x}_0)}\left[-e^{-k\mathcal{E}_0(\mathbf{x}_0)} \log \frac{e^{-f_\phi(\mathbf{x}_t,t)}}{\sum e^{-f_\phi(\mathbf{x}_t,t)}}\right].$$

It has been proved in Lu et al. (2023) that the optimal $f_{\phi^*}(\mathbf{x}_t, t)$ satisfies $\nabla_{\mathbf{x}_t} f_{\phi^*}(\mathbf{x}_t, t) = k\nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t)$, so that sampling can be carried out according to the reverse-time SDE with score function:

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) = s_\theta(\mathbf{x}_t, t) - \eta\nabla_{\mathbf{x}_t} f_\phi(\mathbf{x}_t, t),$$

where η is the hyperparameter controlling the guidance strength, and $s_\theta(\mathbf{x}_t, t)$ refers to the baseline score model in Section 4.1. Note that $s_\theta(\mathbf{x}_t, t)$ and the intermediate energy network $f_\phi(\mathbf{x}_t, t)$ are trained independently.

Since our score network only samples protein backbone structures, we use `faspr` (Huang et al., 2020) to generate full-atom structures for energy evaluation using `OpenMM` (Eastman et al., 2017). See implementation details in Appendix C.3. However, we come across inherent challenges with normalization intricacies in the energy computed by `OpenMM`. Within limited batch size, we witness significant numerical volatility in estimating. In contrast, the numerical value of atomic force will be much more stable. Consequently, the design of a force-guided policy tailored for protein conformation generation proves to be incredibly significant. Compared with energy guidance, force guidance does not suffer from the energy fluctuation problem and can be directly applied as guidance in the reverse sampling process. In the next section, we will illustrate how to realize intermediate force guidance.

4.3 Intermediate Force-Guided Diffusion

In the context of protein conformation modeling, we have access to both a physics-based energy function $\mathcal{E}_0(\mathbf{x}_0)$ (i.e., interatomic potential energy) as well as its gradient $\nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0)$ (i.e., force on each atom). In contrast to the unnormalized potential energy function, atomic force is more local and exhibits better numerical stability, which also aligns better with the score matching objective. Following the marginal distribution of energy in Eq. (4), we have the following properties for the force function.

Proposition 2. *Given the presumption of $p_t(\mathbf{x}_t|\mathbf{x}_0) := q_t(\mathbf{x}_t|\mathbf{x}_0)$, $0 < t \leq 1$, the energy function of intermediate state follows Eq. (4). The corresponding intermediate force has the following formula*

$$\nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t) = \frac{\mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[e^{-k\mathcal{E}_0(\mathbf{x}_0)}\zeta(\mathbf{x}_0, \mathbf{x}_t)\right]}{k\mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[e^{-k\mathcal{E}_0(\mathbf{x}_0)}\right]}, \tag{5}$$

where $\zeta(\mathbf{x}_0, \mathbf{x}_t) = \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t) - \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|\mathbf{x}_0)$. Denoting implicit distribution $p_t(\mathbf{x}_t) \propto q_t(\mathbf{x}_t)e^{-\eta k\mathcal{E}_t(\mathbf{x}_t)}$, the score function follows

$$\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t) = \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t) - \eta\nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t), \tag{6}$$

where η is a hyperparameter controlling the strength of force guidance.

Detailed derivations are provided in Appendix A.

Proposition 2 unveils the precise force at time t, thereby advancing our understanding of the intermediate energy function. The intermediate force formula consists of the ground truth potential energy, the diffusion transition kernel, and a marginal score function (estimated by the score model in Section 4.1). As t approaches 0, $q_t(\mathbf{x}_0|\mathbf{x}_t)$ converges to the delta function and the intermediate force $\nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t)$ converges to the ground truth force. On the other hand, when t approaches 1, $q_t(\mathbf{x}_t|\mathbf{x}_0)$ reduces to $q_t(\mathbf{x}_t)$, and the force $\nabla_{\mathbf{x}_1}\mathcal{E}_1(\mathbf{x}_1)$ vanishes to 0. A comprehensive illustration is provided in Appendix A.

Algorithm 1 Force-guided CONFDIFF

Input: generated data $\mathbf{x}_0$ in a batch K, score model $s_\theta(\mathbf{x}_t, t)$ in Section 4.1, energy $\mathcal{E}_0(\mathbf{x}_0)$, force $\nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0)$, intermediate force network $h_\psi(\mathbf{x}_t, t)$

for training iterations **do**
  $t \sim \mathcal{U}(0, 1)$.
  $\mathbf{x}_t =$ Forward diffusion in Eq. (1).
  $q_t(\mathbf{x}_t|\mathbf{x}_0) \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)I)$
  $\zeta(\mathbf{x}_0, \mathbf{x}_t) = s_\theta(\mathbf{x}_t, t) - \nabla\log q_t(\mathbf{x}_t|\mathbf{x}_0)$.
  $Y = \sum_{\mathbf{x}_0} q_t(\mathbf{x}_t|\mathbf{x}_0)e^{-k\mathcal{E}_0(\mathbf{x}_0)}$.
  $\mathcal{L} = \frac{1}{K}\sum \|h_\psi(\mathbf{x}_t, t) - e^{-k\mathcal{E}_0(\mathbf{x}_0)}\zeta(\mathbf{x}_0, \mathbf{x}_t)/Y\|_2^2$.
  $\min_\psi \mathcal{L}$.
end for

Training process. We employ our baseline score model introduced in Section 4.1 to generate protein conformations from $q_0(\mathbf{x}_0)$, which are then used to train an independent intermediate force network $h_\psi(\mathbf{x}_t, t) : \mathcal{M} \times (0, 1] \to \text{Tan}_{\mathbf{x}_t}\mathcal{M}$, the tangent space of $\mathcal{M}$ at $\mathbf{x}_t$.

Let K be a positive integer denoting the training batch size. To ensure that data in a batch adheres to the same Boltzmann distribution, we first randomly choose a protein sequence, then subsequently draw K samples from this sequence. By adding noise to the sampled data $\mathbf{x}_0$ following the SDE in Eq. (1), we obtain perturbed data at time t as $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon_t, \epsilon_t \sim \mathcal{N}(0, I)$, such that $q_t(\mathbf{x}_t|\mathbf{x}_0) \propto 1/\sqrt{1-\alpha_t}\exp\left(-\|\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0\|^2/2\right)$. We define the intermediate force loss function $\mathcal{L}_{\text{f}}(\psi)$ as

$$\mathbb{E}_{p(t),\mathbf{x}_0,\mathbf{x}_t}\left[\left\|h_\psi(\mathbf{x}_t, t) - \frac{e^{-k\mathcal{E}_0(\mathbf{x}_0)}\zeta(\mathbf{x}_0, \mathbf{x}_t)}{\mathbb{E}_{\mathbf{x}_0^K}\left[q_t(\mathbf{x}_t|\mathbf{x}_0)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\right]}\right\|_2^2\right], \tag{7}$$

where $\zeta(\mathbf{x}_0, \mathbf{x}_t) = \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t) - \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|\mathbf{x}_0)$. The latter component of Eq. (7) signifies the precise value

Algorithm 2 Inference with force-guided CONFDIFF

$\mathrm{x}_N = [\mathrm{T}_N, \mathrm{R}_N] \sim [\mathcal{N}(0, I_3)^{\otimes N}, \mathcal{TN}_{\mathbf{R}}(0, I)^{\otimes N}]$.
# $\mathcal{T}$: the tangent space
for SDE time step $i = N, ..., 1$ **do**
  $\epsilon \sim [\mathcal{N}(0, I_3)^{\otimes N}, \mathcal{TN}_{\mathbf{R}}(0, I)^{\otimes N}]$.
  $[S_{\mathbf{T}_i}, S_{\mathbf{R}_i}] = \gamma s_\theta(\mathbf{x}_i, i, \mathrm{seq}) + (1-\gamma) s_\theta(\mathbf{x}_i, i)$.
  $H = h_\psi(\mathbf{x}_i, i)$.
  $M_{\mathbf{T}_i} = \mathbf{P}(\frac{1}{2}\beta_i \mathbf{T}_i + \beta_i(S_{\mathbf{T}_i} - \eta H) + \sqrt{\beta_i}\epsilon_{\mathbb{R}^{3N}})$.
  $M_{\mathbf{R}_i} = \frac{\mathrm{d}}{\mathrm{d}i}\sigma_i^2 \cdot S_{\mathbf{R}_i} + \sqrt{\frac{\mathrm{d}}{\mathrm{d}i}\sigma_i^2}\epsilon_{\mathbb{SO}(3)^N}$.
  $\mathbf{x}_{i-1} = \exp_{\mathbf{x}_i}\{[M_{\mathbf{T}_i}, M_{\mathbf{R}_i}]\}$.
end for
$\mathrm{x}_0 = [\mathrm{T}_0, \mathrm{R}_0]$.

of the intermediate force at time t, where $\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)$ is the intractable score function estimated by $s_\theta(\mathbf{x}_t, t)$ in our baseline and $q_t(\mathbf{x}_t|\mathbf{x}_0)$ is the tractable Gaussian distribution. Instead of computing $\mathbb{E}_{\mathbf{x}_0}\left[q_t(\mathbf{x}_t|\mathbf{x}_0)e^{-\mathcal{E}_0(\mathbf{x}_0)}\right]$ in the entire generated data collection, we estimate the expectation with the K samples in a batch (Lu et al., 2023). The force-guided training process is summarized in Algorithm 1.

Taking into account that the exact intermediate force formula in Eq. (5) equals to the MD force-field $\nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0)$ at $t = 0$, and turns 0 when $t = 1$, we construct the network $h_\psi(\mathbf{x}_t, t)$ as the interpolation form following (Máté & Fleuret, 2023):

$$h_\psi(\mathbf{x}_t, t) = (1-t)\nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0) + t(1-t)g_\psi(\mathbf{x}_t, t),$$

where $g_\psi(\mathbf{x}_t, t)$ is a neural network estimating the intermediate term within our interpolation construction. It ensures the boundary conditions at $t \in \{0, 1\}$ and we empirically find that this construction reduces the variance during training. Detailed illustration is shown in Appendix B.2.

Inference process. Starting from the prior distribution in $\mathrm{SE}(3)^N$, the entire inference process is illustrated in Algorithm 2. Score network $[S_{\mathbf{T}_i}, S_{\mathbf{R}_i}]$ is the baseline model mixing the sequence conditional score and unconditional score, as introduced in Section 4.1. Force guidance term $h_\psi(\mathbf{x}_t, t)$ is incorporated to the score function term at every time step during the reverse sampling process, whose guidance strength is controlled by hyperparameter η. To maintain structural stability, force guidance is only applied to the translational components (i.e., α-carbons) in $\mathbb{R}^{3N}$.

5 Experiments

We evaluate the effectiveness of CONFDIFF with different guidance strategies on protein conformation generation tasks. In brief, we demonstrate that 1) classifier-free guidance balances the trade-off between sample quality and diversity; 2) energy and force guidance effectively guide the model to sample conformations with lower energy; 3) combining classifier-free and physical guidance strategies, CONFDIFF outperforms current state-of-the-art models by generating high-quality and diverse samples more truthful to the underlying Boltzmann distribution.

5.1 Experimental Setup

Models. We compare CONFDIFF with two state-of-the-art diffusion-based models for protein conformation generation: EIGENFOLD (Jing et al., 2023) samples protein conformations using harmonic diffusion and is fully conditioned on pretrained sequence representations; STR2STR (Lu et al., 2024) first predicts a folded structure using a protein folding model, then perturbs the structure through forward-and-backward diffusion of an unconditional model (sequence-agnostic). All baseline models are run on their default settings. For CONFDIFF, we follow STR2STR to sample conformations with varying sequencing condition levels (γ). All models are trained on experimental protein structures deposited in the Protein Data Bank (PDB) without using additional MD trajectory data.

Datasets. We apply CONFDIFF on two protein conformation benchmarks for a comprehensive evaluation: 1) **fast-folding proteins** comprise MD simulation data for 12 small proteins with fast folding-unfolding dynamics (Lindorff-Larsen et al., 2021), 2) bovine pancreatic trypsin inhibitor (**BPTI**) contains MD simulation data for BPTI exhibiting five metastable states (Shaw et al., 2010). See Appendix C.1 for more details.

Metrics. We evaluate generated conformations as to their *validity*, *precision*, *diversity* and the *similarity of distributions* when comparing conformation ensembles. We summarize key metrics here and include details in Appendix C.4:

- *Validity* evaluates whether generated conformations comply with basic physics as valid proteins. Similar to Lu et al. (2024), we evaluate if a conformation contains steric *clashes* between any residues or *bond breaks* between adjacent residues, based on the coordinates of α-carbons (CA). Conformations without *clash* and *break* are considered valid structures and we calculate the validity VAL-CA as the fraction of valid conformations in all generated samples.
- *Precision* indicates how accurate the model can recover known conformations of a protein. To evaluate structural similarity, sample conformations are aligned to a reference structure and the root-mean-square deviation (RMSD) of α-carbons is calculated. The smallest RMSD between samples and the reference structure is reported as the *precision*.
- *Diversity* shows the model's ability to generate structurally different conformations. We estimate RMSD between pairs of conformations in samples and report the mean RMSD (termed RMSF) as sample diversity.
- *Similarity between distributions* measures how well generated conformations resemble the samples from MD simulation (i.e., Boltzmann distribution). We follow Lu et al.

Table 1. Results on fast-folding proteins. The results for *classifier-free*, *energy guidance*, and *force guidance* are labeled as CONFDIFF-BASE, CONFDIFF-FORCE, and CONFDIFF-ENERGY. All values are shown as mean/median. Best result for each metric is shown in **bold** and the second best is underlined.

MODELS	JS DISTANCE (↓)				VAL-CA	$\text{RMSE}_{\text{CONTACT}}$	RMSF
	PWD	RG	TIC	TIC-2D	(↑)	(↓)	(Å)
EIGENFOLD	0.53/0.56	0.52/0.55	0.50/0.50	0.64/0.66	0.15/0.08	6.18/6.22	1.6/1.1
STR2STR-SDE	0.34/0.32	0.30/0.24	0.39/0.38	0.56/0.58	**0.97/0.98**	3.68/4.01	7.8/8.0
STR2STR-ODE	0.37/0.38	0.33/0.30	0.40/0.39	0.57/0.59	0.96/0.97	4.14/4.36	6.4/6.3
CONFDIFF-BASE	**0.29/0.27**	**0.25/0.22**	**0.36/0.37**	**0.52/0.52**	0.89/0.91	3.61/3.57	6.1/5.9
CONFDIFF-ENERGY	0.34/0.34	0.31/0.29	0.39/0.40	0.54/0.56	0.97/0.97	3.65/3.80	7.1/6.1
CONFDIFF-FORCE	**0.29/0.27**	0.26/0.24	0.38/0.38	0.54/0.54	**0.97/0.98**	**3.25/3.38**	6.2/5.7

(2024) to use Jensen-Shannon distance (JS) to characterize the similarity between generated samples and reference samples from MD. For distribution estimation, α-carbon coordinates are projected to three types of feature dimensions: pairwise distance between α-carbon atoms (PWD); radius-of-gyration (RG) that is the distance of a α-carbon atom to the center-of-mass, and time-lagged independent components (TIC), a reduced feature space related to protein dynamics. We compute JS distance on each feature dimension and report the mean distance; for TIC, we also report the JS distance between joint distributions projected to the first two TIC components (TIC-2D).

5.2 Conformation Sampling with Force Guidance

As an initial proof-of-concept, we first show that proposed *force guidance* can improve the sampling capability of CONFDIFF by generating protein conformations with reduced energy (i.e. improved stability). We select WW-domain, one of the fast-folding proteins from Lindorff-Larsen et al. (2021) as an example and generate conformations using CONFDIFF at varying degrees of force guidance (η) and sequence condition (γ). As shown in Figure 2, without force guidance ($\eta = 0$), models with weaker sequence condition can generate samples with higher diversity (RMSF), however, also with higher energy, showing the quality-diversity compromise. After integrating force guidance, CONFDIFF mitigates such effect and are able to generates samples with lower energy while maintaining similar RMSF levels. This result shows that *force guidance* can improve conformation stability without drastically decease the diversity. This favorable effect suggests integrating physical information might provide fine-grained guidance to sample locally optimized structures.

5.3 Distribution Prediction for Fast-Folding Proteins

This dataset consists of 12 short proteins whose folding and unfolding dynamics have been examined in long-time all-atom MD simulations (Lindorff-Larsen et al., 2021). In this experiment, we evaluate the capability of models to recover the conformation distributions observed in the simulation.

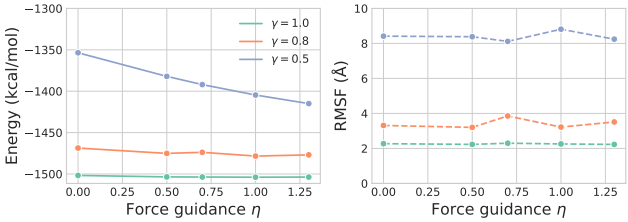

Figure 2. Energy (left) and diversity (right) of sampled conformations for WW-domain with various levels of force guidance (η) and sequence condition (γ). Models with weaker sequence condition generate more diverse samples and *force guidance* improves conformation stability without drastically decease the diversity.

This benchmark requires the model to not only predict the stable folded structures, but also physically viable unfolded or partially unfolded structures. To assess distribution similarity, we evaluate JS-distances between the distributions of generated samples and ground truth MD samples, along with the *validity* score and RMSF for *diversity*. In addition, we report the root-mean-square-error on the predicted contact rate between residues $\text{RMSE}_{\text{contact}}$ to reflect the model accuracy in flexible regions (Janson et al., 2023).

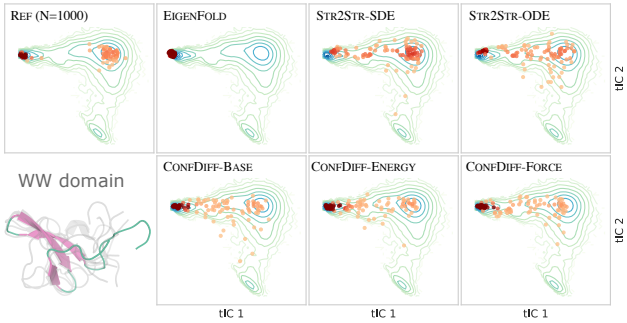

Figure 3. Sample distribution over the first two TIC components for WW-domain. REF (N=1000) shows 1000 random samples from the reference MD simulation. The illustration in the lower left shows the experimental structure in the folded state (in color) and 5 random samples from the reference (in grey).

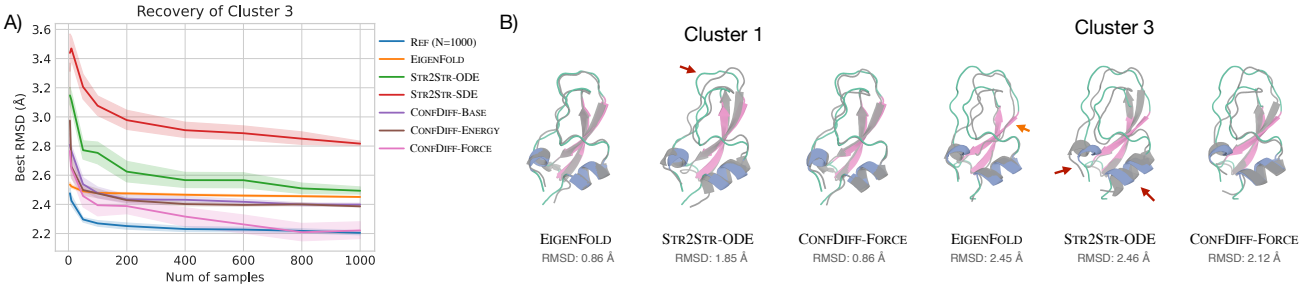

Figure 4. Metastable state prediction for BPTI. A) The precision of predicting Cluster 3 vs varying sample sizes. B) Visual comparison of best samples of three models for Cluster 1 and Cluster 3. Reference structures are shown in color and sample structures are in grey. RMSD vs reference is labeled. For Cluster 1, STR2STR shows lower accuracy in the upper loop region (red arrow); for Cluster 3, the more challenging task, EIGENFOLD does not correctly predict the β-sheet (orange arrow) and STR2STR shows more global structural misalignment (red arrows) that might contribute to higher RMSD.

We summarize results in Table 1 and showcase the sample distribution in TIC projection in Figure 3 and Figure S4. CONFDIFF consistently outperforms EIGENFOLD and STR2STR in terms of recovering sample distributions with lower JS distance and lower RMSE in predicting residue-residue contacts. Notably, integrating energy and force guidance improves the sample validity over CONFDIFF-Base and sustains similar sample diversity, confirming the benefits of integrating physical guidance.

In comparison, EIGENFOLD shows limited sample diversity likely due to the strong tendency to predict the folded state as a fully-conditional diffusion model. STR2STR leverages the unconditional diffusion model to explore diverse structural space and generate diverse samples. Nevertheless, better distribution-related scores of CONFDIFF suggests diffusion process controlled by sequence condition might be generate samples more truthful to a protein's true distribution.

5.4 Metastable States Prediction for BPTI

Previous MD simulation studies on BPTI have recovered five kinetic clusters, representing metastable states near its native folded state (Shaw et al., 2010). For this benchmark, we evaluate the quality and efficiency for models to recover these five states. Compared with fast-folding proteins, this benchmark requires model to generate different conformations while maintaining the correct folding structure.

We measure model *precision* by the lowest sample RMSD to cluster centers. For comparison, we report the best RMSD averaged over five clusters (RMSDAVG) and the RMSD to cluster 3 (RMSDCLS3), the cluster that is the most difficult to sample. As shown in Table 2, CONFDIFF with force guidance performs the best on both metrics, suggesting better capability on predicting distinct metastable states.

We further assess the model efficiency by comparing the precision at varying sample sizes. As shown in Figure 4A, all CONFDIFF models show good efficiency to sample the remote cluster 3 while EIGENFOLD quickly plateaued due to its low sample diversity. CONFDIFF-Force showed best performance, and even close to REF (N=1000), 1000 conformations randomly sampled from MD trajectories.

Table 2. Recovery of five metastable states for BPTI. RMSDAVG denotes the best RMSD score, averaged over five states. RMSDCLS3 refers to the best RMSD to the remote cluster 3, which is the most difficult to sample. (Unit: Å.)

MODEL	RMSDAVG (↓)	RMSDCLS3 (↓)
REFERENCE	1.10	2.19
EIGENFOLD	1.47	2.45
STR2STR-ODE	2.06	2.46
STR2STR-SDE	2.26	2.80
CONFDIFF-BASE	1.55	2.38
CONFDIFF-ENERGY	1.57	2.38
CONFDIFF-FORCE	**1.41**	**2.12**

These results show superb performance of CONFDIFF models, particularly CONFDIFF-FORCE in predicting distinct metastable states with high precision and potential to sample remote states.

6 Conclusion and Future Work

In this work, we introduce CONFDIFF for protein conformation generation. By mixing a sequence-conditional score network with an unconditional model, we strike a balance between sampling quality and diversity through classifier-free guidance. Upon such a baseline model, we additionally propose novel physics-based energy and force guidance strategies with theoretical guarantee, which effectively guide the diffusion sampler to generate low energy conformations better complying with the underlying Boltzmann distribution. Experiments on a variety of protein systems demonstrate the effectiveness of our proposed method.

Although CONFDIFF has shown promising performance, there is still room for further improvement. Due to limited conformation diversity in existing protein structure databases, models fully conditioned on sequence informa-

tion tend to predict the folded structure and lack the capability to fully picture the entire conformation landscape. On the other hand, the full-atom MD energy function is still relatively computationally intensive and time-consuming to use as a preference function under the generative modeling framework. Nevertheless, it is our first attempt to combine the Boltzmann prior with diffusion models and we will persist in the pursuit of more efficient approaches for force guidance, as well as effective measures to minimize the variance during training.

Acknowledgements

We would like to express our sincere appreciation to Dr. Hang Li for his invaluable support. Concurrently, we would like to extend our gratitude to Yi Zhou for his insightful comments and guidance on this project, which provided much inspiration.

Impact Statement

Our work on protein conformation can be used to generate diverse and high quality structures, and may be adapted to other scenarios of computer-aided design, such as reinforce learning, small molecule design, and antibody design. It is imperative to affirm the responsible application of our approach and emphasize a commitment to refrain from employing it for any detrimental or malicious purposes.

References

Anand, N. and Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

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

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

Bose, A. J., Akhound-Sadegh, T., Fatras, K., Huguet, G., Rector-Brooks, J., Liu, C.-H., Nica, A. C., Korablyov, M., Bronstein, M., and Tong, A. Se (3)-stochastic flow matching for protein backbone generation. *arXiv preprint arXiv:2310.02391*, 2023.

De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y. W., and Doucet, A. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022.

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

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

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7): e1005659, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

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*, 2022.

Hoffmann, M., Scherer, M. K., Hempel, T., Mardt, A., de Silva, B., Husic, B. E., Klus, S., Wu, H., Kutz, J. N., Brunton, S., and Noé, F. Deeptime: a python library for machine learning dynamical models from time series data. *Machine Learning: Science and Technology*, 2021.

Huang, X., Pearce, R., and Zhang, Y. FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics*, 36(12):3758–3765, 04 2020. doi: 10.1093/bioinformatics/btaa234. URL https://doi.org/10.1093/bioinformatics/btaa234.

Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., Frappier, V., Lord, D. M., Ng-Thow-Hing, C., Van Vlack, E. R., et al. Illuminating protein space with a programmable generative model. *Nature*, pp. 1–9, 2023.

Janson, G., Valdes-Garcia, G., Heo, L., and Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nature Communications*, 14(1):774, 2023.

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.

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

Leach, A., Schmon, S. M., Degiacomi, M. T., and Willcocks, C. G. Denoising diffusion probabilistic models on so (3) for rotational alignment. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2021. doi: 10.1126/science.1208351. URL https://www.science.org/doi/10.1126/science.1208351.

Lu, C., Chen, H., Chen, J., Su, H., Li, C., and Zhu, J. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. *arXiv preprint arXiv:2304.12824*, 2023.

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*, 2024.

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., and Ma, J. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35: 9754–9767, 2022.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015. doi: 10.1021/acs.jctc.5b00255. PMID: 26574453.

Mansoor, S., Baek, M., Park, H., Lee, G. R., and Baker, D. Protein ensemble generation through variational autoencoder latent space sampling. *bioRxiv*, pp. 2023–08, 2023.

Martinkus, K., Ludwiczak, J., Cho, K., Lian, W.-C., Lafrance-Vanasse, J., Hotzel, I., Rajpal, A., Wu, Y., Bonneau, R., Gligorijevic, V., et al. Abdiffuser: Full-atom generation of in-vitro functioning antibodies. *arXiv preprint arXiv:2308.05027*, 2023.

Máté, B. and Fleuret, F. Learning interpolations between boltzmann densities. *Transactions on Machine Learning Research*, 2023.

Naritomi, Y. and Fuchigami, S. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *The Journal of Chemical Physics*, 139(21):215102, 12 2013. ISSN 0021-9606. doi: 10.1063/1.4834695. URL https://doi.org/10.1063/1.4834695.

Nguyen, H., Roe, D. R., and Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *Journal of Chemical Theory and Computation*, 9(4):2020–2034, 2013. doi: 10.1021/ct3010485. PMID: 25788871.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

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.

Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G., and Noé, F. Identification of slow molecular order parameters for markov model construction. *The Journal of chemical physics*, 139(1), 2013.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Saldaño, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Velez Rueda, A. J., Gonik, E., García Melani, A., Novomisky Nechcoff, J., Salas, M. N., et al. Impact of protein conformational diversity on alphafold predictions. *Bioinformatics*, 38(10):2742–2748, 2022.

Scherer, M. K., Trendelkamp-Schroer, B., Paul, F., Pérez-Hernández, G., Hoffmann, M., Plattner, N., Wehmeyer, C., Prinz, J.-H., and Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11:5525–5542, October 2015. ISSN 1549-9618. doi: 10.1021/acs.jctc.5b00743. URL http://dx.doi.org/10.1021/acs.jctc.5b00743.

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010. doi: 10.1126/science.1187409. URL https://www.science.org/doi/abs/10.1126/science.1187409.

Shaw, D. E., Adams, P. J., Azaria, A., Bank, J. A., Batson, B., Bell, A., Bergdorf, M., Bhatt, J., Butts, J. A., Correia, T., et al. Anton 3: twenty microseconds of molecular dynamics simulation before lunch. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–11, 2021.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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

Trendelkamp-Schroer, B. and Noé, F. Efficient estimation of rare-event kinetics. *Physical Review X*, 6(1):011009, 2016.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Vani, B. P., Aranganathan, A., Wang, D., and Tiwary, P. From sequence to boltzmann weighted ensemble of structures with alphafold2-rave. *bioRxiv*, pp. 2022–05, 2022.

Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M. K., Greenwood, J., et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7): 2695–2703, 2015.

Wang, Y., Guo, L., Wu, H., and Zhou, T. Energy based diffusion generator for efficient sampling of boltzmann distributions. *arXiv preprint arXiv:2401.02080*, 2024.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *BioRxiv*, pp. 2022–12, 2022.

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*, pp. 1–3, 2023.

Wu, K. E., Yang, K. K., Berg, R. v. d., Zou, J. Y., Lu, A. X., and Amini, A. P. Protein structure generation via folding diffusion. *arXiv preprint arXiv:2209.15611*, 2022a.

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

Yim, J., Campbell, A., Foong, A. Y., Gastegger, M., Jiménez-Luna, J., Lewis, S., Satorras, V. G., Veeling, B. S., Barzilay, R., Jaakkola, T., et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023a.

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. *arXiv preprint arXiv:2302.02277*, 2023b.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702—710, December 2004. ISSN 0887-3585. doi: 10.1002/prot.20264.

Zheng, S., He, J., Liu, C., Shi, Y., Lu, Z., Feng, W., Ju, F., Wang, J., Zhu, J., Min, Y., et al. Towards predicting equilibrium distributions for molecular systems with deep learning. *arXiv preprint arXiv:2306.05445*, 2023.

A Proof of Propositions 1 and 2

Proof. Given $p_0(\mathbf{x}_0) = q_0(\mathbf{x}_0)e^{-k\mathcal{E}(\mathbf{x}_0)}/Z$ and $p_t(\mathbf{x}_t|\mathbf{x}_0) = q_t(\mathbf{x}_t|\mathbf{x}_0)$, the marginal distribution $p_t(\mathbf{x}_t)$ follows

$$\begin{aligned} p_t(\mathbf{x}_t) &= \int p_t(\mathbf{x}_t|\mathbf{x}_0)p_0(\mathbf{x}_0)\ \mathrm{d}\mathbf{x}_0 \\ &= \int q_t(\mathbf{x}_t|\mathbf{x}_0)\frac{q_0(\mathbf{x}_0)e^{-k\mathcal{E}(\mathbf{x}_0)}}{Z}\ \mathrm{d}\mathbf{x}_0 \\ &= \int q_t(\mathbf{x}_0|\mathbf{x}_t)q_t(\mathbf{x}_t)\frac{e^{-k\mathcal{E}(\mathbf{x}_0)}}{Z}\ \mathrm{d}\mathbf{x}_0 \\ &= q_t(\mathbf{x}_t)\mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[\frac{e^{-k\mathcal{E}(\mathbf{x}_0)}}{Z}\right]. \end{aligned}$$

Under the assumption $p_t(\mathbf{x}_t) := q_t(\mathbf{x}_t)e^{-k\mathcal{E}_t(\mathbf{x}_t)}$, $\mathcal{E}_t(\mathbf{x}_t)$ has the formula

$$\mathcal{E}_t(\mathbf{x}_t) = \begin{cases} \mathcal{E}_0(\mathbf{x}_0), & t = 0 \\ -\frac{1}{k}\log \mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[e^{-k\mathcal{E}(\mathbf{x}_0)}\right] + \log Z, & 0 < t \leq 1 \end{cases}.$$

Considering that the constant coefficient has no influence on the computation of the force field, we thus propose a notation where $\mathcal{E}_t(\mathbf{x}_t) := -\frac{1}{k}\log \mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[e^{-k\mathcal{E}(\mathbf{x}_0)}\right]$, $0 < t \leq 1$.
Given energy function $\mathcal{E}_t(\mathbf{x}_t)$ at time t, the force equation of intermediary states can be derived by

$$\nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t) = -\frac{\int e^{-k\mathcal{E}_0(\mathbf{x}_0)}\nabla_{\mathbf{x}_t}q_t(\mathbf{x}_0|\mathbf{x}_t)\ \mathrm{d}\mathbf{x}_0}{k\int q_t(\mathbf{x}_0|\mathbf{x}_t)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\ \mathrm{d}\mathbf{x}_0},$$

where the numerator adopt

$$\begin{aligned} \int e^{-k\mathcal{E}_0(\mathbf{x}_0)}\nabla_{\mathbf{x}_t}q_t(\mathbf{x}_0|\mathbf{x}_t)\ \mathrm{d}\mathbf{x}_0 &= \int e^{-k\mathcal{E}_0(\mathbf{x}_0)}q_t(\mathbf{x}_0|\mathbf{x}_t)\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_0|\mathbf{x}_t)\ \mathrm{d}\mathbf{x}_0 \\ &= \int q_t(\mathbf{x}_0|\mathbf{x}_t)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\nabla_{\mathbf{x}_t}\log\frac{q_t(\mathbf{x}_t|\mathbf{x}_0)q_0(\mathbf{x}_0)}{q_t(\mathbf{x}_t)}\ \mathrm{d}\mathbf{x}_0 \\ &= \int q_t(\mathbf{x}_0|\mathbf{x}_t)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\left(\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|\mathbf{x}_0) - \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)\right)\ \mathrm{d}\mathbf{x}_0 \\ &= \mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[e^{-k\mathcal{E}_0(\mathbf{x}_0)}\left(\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|\mathbf{x}_0) - \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)\right)\right]. \end{aligned}$$

We define $\zeta(\mathbf{x}_0, \mathbf{x}_t) := \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t) - \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|\mathbf{x}_0)$. Therefore

$$\begin{aligned} \nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t) &= -\frac{\mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[-e^{-k\mathcal{E}_0(\mathbf{x}_0)}\zeta(\mathbf{x}_0, \mathbf{x}_t)\right]}{k\mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[e^{-k\mathcal{E}_0(\mathbf{x}_0)}\right]} \\ &= \frac{\mathbb{E}_{q_0(\mathbf{x}_0)}\left[q_t(\mathbf{x}_t|\mathbf{x}_0)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\zeta(\mathbf{x}_0, \mathbf{x}_t)\right]}{k\mathbb{E}_{q_0(\mathbf{x}_0)}\left[q_t(\mathbf{x}_t|\mathbf{x}_0)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\right]}. \end{aligned}$$

As t approaches 0, it prompts $q_t(\mathbf{x}_0|\mathbf{x}_t)$ involving into $\delta(\mathbf{x}_t - \mathbf{x}_0)$. Subsequently, the intermediate force turns into

$$\begin{aligned} \nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t) &= -\frac{\int e^{-k\mathcal{E}_0(\mathbf{x}_0)}\nabla_{\mathbf{x}_t}q_t(\mathbf{x}_0|\mathbf{x}_t)\ \mathrm{d}\mathbf{x}_0}{k\int q_t(\mathbf{x}_0|\mathbf{x}_t)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\ \mathrm{d}\mathbf{x}_0} \\ &= -\frac{\mathbb{E}_{q_t(\mathbf{x}_0|\mathbf{x}_t)}\left[e^{-k\mathcal{E}_0(\mathbf{x}_0)}/q_t(\mathbf{x}_0|\mathbf{x}_t)\cdot\nabla_{\mathbf{x}_t}\delta(\mathbf{x}_t - \mathbf{x}_0)\right]}{k\int \delta(\mathbf{x}_t - \mathbf{x}_0)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\ \mathrm{d}\mathbf{x}_0} \\ &= -\frac{\nabla_{\mathbf{x}_t}\left[e^{-k\mathcal{E}_0(\mathbf{x}_0)}/q_t(\mathbf{x}_0|\mathbf{x}_t)\cdot q_t(\mathbf{x}_0|\mathbf{x}_t)\right]}{ke^{-k\mathcal{E}_0(\mathbf{x}_0)}} \\ &= \nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0). \end{aligned}$$

As t approaches 1, both $p_t(\mathbf{x}_t)$ and $q_t(\mathbf{x}_t)$ become Gaussian distribution, prompting the energy $e^{-k\mathcal{E}_t(\mathbf{x}_t)}$ becoming a constant and force $\nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t)$ becoming 0. This observation is in alignment with the theoretical deductions, as when $t \to 1$, $q_t(\mathbf{x}_t|\mathbf{x}_0) \to q_t(\mathbf{x}_t)$ and $\zeta(\mathbf{x}_0, \mathbf{x}_t) \to 0$, thus $\nabla_{\mathbf{x}_t}\mathcal{E}_t(\mathbf{x}_t)$ becomes 0. □

B CONFDIFF Architecture

B.1 Diffusion on SE(3)

Diffusion models on $\mathrm{SE}(3)^N$ demonstrates powerful generative capability and are widely applied in the field of protein backbone design (De Bortoli et al., 2022; Yim et al., 2023b). Conducted independently over $\mathbb{R}^{3N}$ and $\mathrm{SO}(3)^N$, the diffusion process over position variables $\{\mathbf{x}_t\}_{t\in[0,1]} = [\mathbf{T}_t, \mathbf{R}_t]_{t\in[0,1]}$ are defined as

$$\mathrm{d}\mathbf{x}_t = \left[-\frac{1}{2}\beta_t \mathbf{P}\mathbf{T}_t, 0\right]\mathrm{d}t + \left[\sqrt{\beta_t}\mathbf{P}\mathrm{d}\mathbf{w}_t, \sqrt{\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2}\mathrm{d}\mathbf{w}_t^{\mathrm{SO}(3)}\right], \quad \mathbf{x}_0 \sim q_0(\mathbf{x}_0)$$

where $[\mathbf{w}_t, \mathbf{w}_t^{\mathrm{SO}(3)}]$ is the standard Wiener process in $[\mathcal{N}(0, I_3)^{\otimes N}, \mathcal{U}(\mathrm{SO}(3))^{\otimes N}]$. Here, $\beta_t = \beta_{\min} + t(\beta_{\max} - \beta_{\min})$ and we set $\beta_{\min} = 0.1$, $\beta_{\max} = 20$. Similarly, $\sigma(t) = \log(te^{\sigma_{\max}} + (1-t)e^{\sigma_{\min}})$ and we set $\sigma_{\min} = 0.1$, $\sigma_{\max} = 1.5$. Adding noise according to the aforementioned equation, the transition kernel satisfies $q_t(\mathbf{T}_t|\mathbf{T}_0) = \mathcal{N}(\mathbf{T}_t; \sqrt{\alpha_t}\mathbf{T}_0, (1-\alpha_t)I)$, where $\alpha_t := \exp\left(-\int_0^t \beta(s)\,\mathrm{d}s\right)$, and the rotation kernel satisfies $p_t(\mathbf{R}_t|\mathbf{R}_0) = \mathrm{IGSO}_3(\mathbf{R}_t; \mathbf{R}_0, \sigma_t^2)$,

$$\mathrm{IGSO}_3(\mathbf{R}_t; \mathbf{R}_0, \sigma_t^2) = \sum_{\ell}(2\ell+1)\mathrm{e}^{-\frac{\ell(\ell+1)\sigma_t^2}{2}}\frac{\sin((\ell+\frac{1}{2})\omega_t)}{\sin(\frac{\omega_t}{2})},$$

where $\omega_t = \mathrm{Axis_angle}(\mathbf{R}_0^\top \mathbf{R}_t)$. Correspondingly, the associated reverse time diffusion process follows

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{P}\left(-\frac{1}{2}\beta_t\mathbf{T}_t - \beta_t\nabla\log q_t(\mathbf{T}_t)\right), -\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2\nabla\log q_t(\mathbf{R}_t)\right]\mathrm{d}t + \left[\sqrt{\beta_t}\mathbf{P}\mathrm{d}\bar{\mathbf{w}}_t, \sqrt{\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2}\mathrm{d}\bar{\mathbf{w}}_t^{\mathrm{SO}(3)}\right].$$

Considering the intractable marginal score function $\nabla\log q_t(\mathbf{x}_t) = [\nabla\log q_t(\mathbf{T}_t), \nabla\log q_t(\mathbf{R}_t)]$, we use score matching techniques to estimate them respectively. Training the neural networks $s_\theta(\mathbf{T}_t, t)$ and $s_\theta(\mathbf{R}_t, t)$, the denoising score matching loss function is defined as

$$\begin{aligned}
\mathcal{L}_{\mathrm{DSM}}^{\mathbf{T}} &= \mathbb{E}_{p(t)}\mathbb{E}_{q_0(\mathbf{T}_0), q_t(\mathbf{T}_t|\mathbf{T}_0)}\left[\lambda(t)\|s_\theta(\mathbf{T}_t, t) - \nabla_{\mathbf{T}_t}\log p_t(\mathbf{T}_t|\mathbf{T}_0)\|^2\right],\\
\mathcal{L}_{\mathrm{DSM}}^{\mathbf{R}} &= \mathbb{E}_{p(t)}\mathbb{E}_{q_0(\mathbf{R}_0), q_t(\mathbf{R}_t|\mathbf{R}_0)}\left[\lambda^r(t)\|s_\theta^r(\mathbf{R}_t, t) - \nabla_{\mathbf{R}_t}\log p_t(\mathbf{R}_t|\mathbf{R}_0)\|^2\right],
\end{aligned}$$

where $\lambda(t)$, $\lambda^r(t)$ is a time weighting function and the conditional score has the following formula

$$\begin{aligned}
\nabla_{\mathbf{T}_t}\log p_t(\mathbf{T}_t|\mathbf{T}_0) &= -\frac{\mathbf{T}_t - \sqrt{\alpha_t}\mathbf{T}_0}{1-\alpha_t},\\
\nabla_{\mathbf{R}_t}\log p_t(\mathbf{R}_t|\mathbf{R}_0) &= \frac{\mathbf{R}_t}{\omega_t}\log\left(\mathbf{R}_0^\top\mathbf{R}_t\right)\partial_{\omega_t}\log\mathrm{IGSO}_3\left(\mathbf{R}_t; \mathbf{R}_0, \sigma_t^2\right).
\end{aligned}$$

To improve the quality of local atomic structures, auxiliary losses are added to the optimization problem, including the MSE loss on the backbone (bb) positions $\mathcal{L}_{\mathrm{bb}}$ and a local neighborhood loss on pairwise atomic distances $\mathcal{L}_{\mathrm{2D}}$. The total loss can be written as

$$\mathcal{L} = \mathcal{L}_{\mathrm{DSM}}^{\mathbf{T}} + 0.5 * \mathcal{L}_{\mathrm{DSM}}^{\mathbf{R}} + 0.25 * 1\left\{t < \frac{\mathrm{T_F} = 0.25}{4}\right\}(\mathcal{L}_{\mathrm{bb}} + \mathcal{L}_{\mathrm{2D}}).$$

B.2 Interpolation with Simulation Data

According to the derivation in Appendix A, $\nabla_{\mathrm{x}_0}\mathcal{E}_0(\mathrm{x}_0)$ has the ground truth value at the boundary $t \in \{0, 1\}$ and we observe that both score functions $\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|\mathbf{x}_0)$ and $\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)$ exhibit large variance at $t = 0$ (due to small noise level at small t). Thus, we propose an interpolation strategy for the intermediate force function like in (Máté & Fleuret, 2023; Wang et al., 2024):

$$h_\psi(\mathbf{x}_t, t) = (1-t)\nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0) + t(1-t)g_\psi(\mathbf{x}_t, t).$$

where $g_\psi(\mathbf{x}_t, t)$ is a neural network to estimate the intermediate term of interpolation construction. It ensures that as t approaches 0, the network approximates the ground truth force $\nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0)$ and as t nears 1, the network becomes 0.

$$h_\psi(\mathbf{x}_t, t) = \begin{cases}\nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0), & t = 0\\ \text{Parametrized Interpolation}, & 0 < t < 1\\ 0, & t = 1\end{cases}.$$

Nevertheless, throughout the inference process, despite the ability to derive a predicted $\widehat{\mathbf{x}}_0$ at every step for the computation of $\nabla_{\widehat{\mathbf{x}}_0}\mathcal{E}_0(\widehat{\mathbf{x}}_0)$, the utilization of OpenMM for energy calculation is highly time-consuming, particularly in estimating energy in lengthy protein chains. To further demonstrate the computational cost of potential energy evaluation, we tabulate the total time consumption of once energy/force calculation using openMM (the cost of side-chain packing using faspr is excluded) for the fast-folding proteins. Due to the 1000 steps in the inference diffusion process, the time required to generate a single protein conformation is multiplied by 1000.

Table S1. The time consumption of once energy/force calculation using openMM.

	1FME	2F4K	2JOF	2WAV	A3D	CLN025
LENGTH	28	35	20	47	73	10
CPU	3.6±0.6 s	4.2±1.8 s	1.4±0.3 s	8.3±3.3 s	19.4±5.3 s	0.9±0.2 s
GPU	1.7±0.2 s	1.7±0.5 s	1.6±0.5 s	1.9±0.6 s	2.2±0.5 s	1.6±0.5 s
	GTT	NTL9	NUG2	PRB	UVF	LAMBDA
LENGTH	35	39	56	47	52	80
CPU	5.0±2.1 s	6.8±1.4 s	12.1±5.5 s	8.7±1.6 s	12.6±2.5 s	26.6±16.4 s
GPU	1.8±0.5 s	1.8±0.6 s	2.0±0.5 s	1.8±0.5 s	2.0±0.5 s	2.0±0.5 s

Considering that the reverse diffusion process is discretized into 1000 steps, the time expenditure is not feasible to incorporate the ground-truth MD force field in each reverse sampling step, necessitating the employment of a surrogate model. Thus, we additionally train a network $\widetilde{g}_\nu(\mathbf{x}_t, t)$ to estimate the ground truth force field $\nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0)$ at the initial moment by MSE loss function $\mathcal{L}_{t=0}(\nu)$, resulting in the transformation of $h_\psi(\mathbf{x}_t, t)$ into

$$h_{\psi,\nu}(\mathbf{x}_t, t) = (1-t)\widetilde{g}_\nu(\mathbf{x}_t, t) + t(1-t)g_\psi(\mathbf{x}_t, t). \tag{8}$$

The optimization problem, combining two loss functions, is defined as:

$$\min_{\psi,\nu} \mathcal{L}(\psi,\nu) = \min_{\psi,\nu} \left[\mathcal{L}_{\text{force}}(\psi,\nu) + \mathcal{L}_{t=0}(\nu)\right],$$

where

$$\mathcal{L}_{\text{force}}(\psi,\nu) = \mathbb{E}_{p(t),\mathbf{x}_0,\mathbf{x}_t}\left[\left\| h_{\psi,\nu}(\mathbf{x}_t, t) - \frac{e^{-k\mathcal{E}_0(\mathbf{x}_0)}\zeta(\mathbf{x}_0, \mathbf{x}_t)}{\mathbb{E}_{\mathbf{x}_0^K}\left[q_t(\mathbf{x}_t|\mathbf{x}_0)e^{-k\mathcal{E}_0(\mathbf{x}_0)}\right]}\right\|_2^2\right]$$
$$\mathcal{L}_{t=0}(\nu) = \mathbb{E}_{p(t)}\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t}\left[\|\widetilde{g}_\nu(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_0}\mathcal{E}_0(\mathbf{x}_0)\|_2^2\right].$$

B.3 Implementation

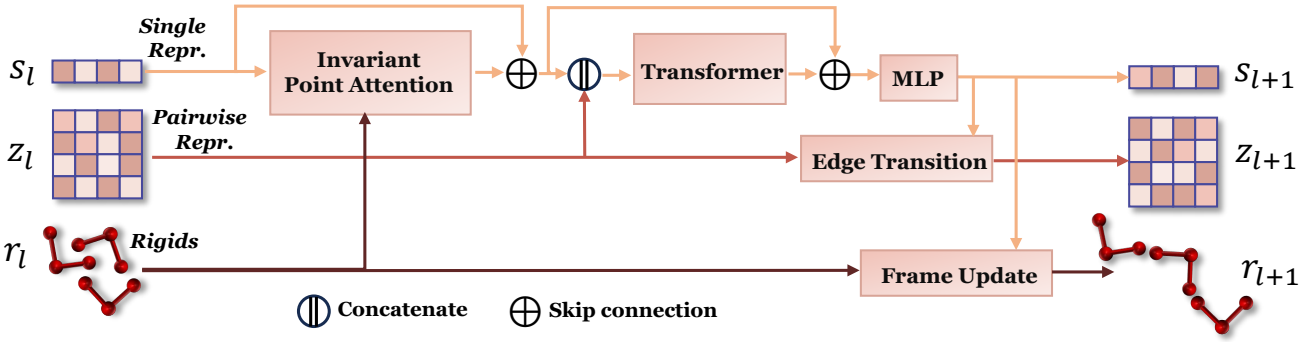

Figure S1. single-layer of CONFDIFF. This architecture, drawing references from Jumper et al. (2021); Yim et al. (2023b), can be widely applied in the unconditional model, the conditional model, the intermediate energy prediction network $f_\phi(\mathbf{x}_t, t)$, the intermediate force prediction network $g_\nu(\mathbf{x}_t, t)$ and the force-field prediction network $\widetilde{g}_\nu(\mathbf{x}_t, t)$ used in Eq. 8. The invariant point attention, edge transition, and frame update modules similar to the corresponding structures in Jumper et al. (2021)

The single layer of CONFDIFF consists of invariant point attention layer (Jumper et al., 2021), transformer and other structure modules similar to Yim et al. (2023b). In the unconditional diffusion model, the single representation s_0 is initialized as a concatenation of the position encoding of residues and sinusoidal time embedding. Meanwhile, the pair representation z_0 is initialized as a concatenation of the relative residue position encoding, sinusoidal time embedding, and relative sequence distances. In the conditional model, the node embedding of ESMfold Lin et al. (2022) with corresponding sequences is additionally concatenated to the single representation s_0. In addition, we apply the same noise-injection schedule as the translation term in Eq. (1) to the ESMFold node embeddings before feeding to the neural networks. This noise injection step further increases sampling diversity. In both the intermediate force prediction network $g_\nu(\mathbf{x}_t, t)$ and the force-field prediction network $\widetilde{g}_\nu(\mathbf{x}_t, t)$, the features are initialized as the output single and pair representation of the conditional model. These features are then used to predict the intermediate force, which is derived from the last three dimensions of the newly predicted rigids. The intermediate energy prediction network $f_\phi(\mathbf{x}_t, t)$ also takes the output of the conditional model as its input and a linear layer is added at the end that transforms the single representation into a scalar value, thereby aiming to predict energy. Detailed hyperparameters are listed in B.5.

B.4 Computational Cost

Our CONFDIFF-base model has similar computational complexity to other diffusion-based conformation generation models, e.g., FrameDiff (Yim et al., 2023b) and Str2Str (Lu et al., 2024). The introduction of a guidance network will not significantly increase computational cost. Specifically, the energy/force guidance modules only introduce a small computational overhead during inference through a **lightweight energy/force prediction head**, which is based on the pretrained CONFDIFF-base model to accomplish guidance.

We made a comparative study of the time and GPU memory requirement for generating protein conformations across different sequence lengths. We used a single NVIDIA-V100 GPU to benchmark model performance. The batch size is 16, and the number of reverse sampling steps is 1000.

Table S2. The time and memory requirement for generating protein conformations across different sequence lengths.

	CONFDIFF-BASE		CONFDIFF-ENERGY		CONFDIFF-FORCE	
LEN (PROTEIN)	TIME (SEC)	MEMORY (MB)	TIME (SEC)	MEMORY (MB)	TIME (SEC)	MEMORY (MB)
58 (BPT)	71	1838	110	3142	104	1916
118 (1JM4.B)	167	3932	256	9722	206	4010
238 (2LAO.A)	578	12402	888	29886	709	12480

Notably, the energy guidance policy almost doubles the memory cost compared to CONFDIFF-base, since the derivative of energy is required to compute the force guidance term. Meanwhile, force guidance only consumes negligibly more computational resources compared to CONFDIFF-base, which is another significant advantage of the force guidance strategy versus the energy guidance strategy.

B.5 Hyperparameters

Table S3. Hyperparameter choices of CONFDIFF

Hyperparameters	Values
Neural network	
Number of IPA blocks	4
Dimension of single repr.	256
Dimension of pairwise Repr.	128
Dimension of hidden	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
SDE	
Number of time steps	1000
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
Training	
Batch Size	32
Learning Rate	1×10^{-4}
Optimizer	Adam (weight decay = 0.01)
Learning Rate Scheduler	ReduceLROnPlateau (factor = 0.5)
Warm up steps	5000
Inference	
Classifier-free guidance strength γ	$[0.5, 0.8, 1.0]$
Force guidance strength η	$[0.5, 0.7, 1.0, 1.3]$

C Experimental Details

C.1 Datasets

Training datasets. We train CONFDIFF using all available single-chain protein structures deposited to the PDB on or before Dec 31, 2021, and perform validation using structures deposited between Jan 1, 2022 and Dec 31, 2022. Unlike previous processing pipelines which typically only extract the first model (i.e., only take "MODEL 1") of each chain, we parse and save all available models in order to make full use of available experimental structures, especially the solution NMR data. We train and validate on protein chains with residue length between 20 and 400, and remove data with less than 70% valid backbone atoms, which leads to 412,371 (57,088) training (validation) structures. During training, we adopt a 30% sequence-similarity-based clustering strategy within our data sampler, where each data loader first randomly samples a cluster, then subsamples a structure from that cluster.

Fast-folding proteins. The simulation details of fast-folding proteins are described in (Lindorff-Larsen et al., 2021) and the trajectories are obtained from the authors. The dataset contains 12 small proteins range from 10 to 80 residues that exhibit multiple fast folding-unfolding events within the millisecond-scale all-atom simulations. Atom coordinates are saved at 200 ps intervals and 520K $\sim$ 14.7M aggregated frames are collected for each protein. α-carbon atom coordinates are used for analysis. For sampling, we generate 1000 conformations for each proteins with guidance strength $\eta = 1.3$ for both *force* and *energy guidance*, and the sequence condition γ sampled between $0.5 \sim 1.0$.

BPTI. Bovine pancreatic trypsin inhibitor (BPTI) is a 58-residue model protein for protein dynamics. Transitions between five meta-stable states are identified in a previous 1 ms simulation study (Shaw et al., 2010). MD trajectories are obtained from the authors and contain atom coordinates saved at 250 ps intervals with 4.1M total frames. α-carbon atom coordinates are used for analysis. The coordinate PDB files for the five kinetic clusters are provided in the supplementary materials for (Shaw et al., 2010). For sampling, we generate 1000 conformations for BPTI with guidance strength $\eta = 1.5$ for *force guidance* and 0.5 for *energy guidance*. The sequence condition γ is set to sample between $0.8 \sim 1.0$.

C.2 Time-Lagged Independent Analysis (TICA)

For fast-folding proteins and BPTI, we use `pyemma` (Scherer et al., 2015) for featurizing protein structures (as pairwise distances between α-carbons), and use `Deeptime` (Hoffmann et al., 2021) for time-lagged independent analysis (TICA) and projecting the protein conformations into the two slowest components for distribution comparison (Naritomi & Fuchigami, 2013; Pérez-Hernández et al., 2013).

C.3 Energy and Force Evaluation

Potential energy of protein structures are evaluated using `OpenMM` (Eastman et al., 2017) following a similar pipeline as in (Lu et al., 2024). First, we use `faspr` (Huang et al., 2020) for side-chain packing (i.e., adding side chain atoms with optimized coordinates) on generated backbone conformations. Protonation (i.e., adding hydrogen atoms), solvation, and energy minimization are performed using `OpenMM` with Amber-ff14SB force field (Maier et al., 2015) and implicit solvent GBn2 (Nguyen et al., 2013). During minimization, we keep the conformations by applying independent harmonic restraints on all heavy (non-hydrogen) atoms with spring constant of 10 $\mathrm{kcal/mol} \cdot \mathrm{\AA}^2$, the tolerance is set to 2.39 $\mathrm{kcal/mol} \cdot \mathrm{\AA}^2$ without maximal step limits. The potential energy excluding the component from harmonic restraints is reported (unit: $\mathrm{kcal/mol}$). Forces on each α-carbon atoms are also reported for training the neural networks used in *force guidance*.

C.4 Details on Evaluation Metrics

Jensen-Shannon (JS) distance between sample distributions. We measure the similarity between two conformation ensembles by JS distance:

$$\mathrm{JS}(p\|q) = \sqrt{\frac{\mathrm{KL}(p\|q) + \mathrm{KL}(q\|p)}{2}},$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence. Similar to (Lu et al., 2024), we project coordinates to three types of feature spaces to compute conformation distribution: 1) Pairwise Distance (PwD) between α-carbon of residues excluding adjacent residues within an offset of 3; 2) Radius of gyrations (Rg) that are distances of α-carbon atoms to the center-of-mass of the conformation; 3) Time-lagged Independent Components (TIC) from TICA analysis on the reference MD trajectories (see Appendix C.2), only the first two components are used for JS calculation. Feature dimensions are discretized with 50 bins according to the values of the reference MD samples and a pseudo count of 10^{-6} is added. We compute the distance over each dimension and report the mean values as the final JS distance.

The first two dimensions of TIC represent the a low dimensional 2D space with features most predictive for protein dynamics, therefore we also estimate the sample distribution of the joint space of these two dimensions, labeled as TIC-2D.

Contact rates between residues. For fast-folding proteins we compute the contact rate between a pair of residues as the fraction of conformation with contacting residues for each pair of residues at a threshold of 10 Å:

$$r(u, v) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{d_i(u, v) \leq 10\text{Å}\},$$

where $d_i(u, v)$ is the Euclidean distance between the α-carbon of residue u and v, N is the number of samples, and $\mathbf{1}\{\cdot\}$ is the indicator function.

To quantify the accuracy of predicting the residue contacts, we compute the root-mean-square error between generated samples and reference MD samples:

$$\text{RMSE}_{\text{contact}} = \sqrt{\frac{2}{L(L-1)} \sum_{u=1}^{L-1} \sum_{v=u+1}^{L} \big(r(u, v) - r_{\text{ref}}(u, v)\big)^2}.$$

Validity. We adopt the *Validity* metrics in (Lu et al., 2024) to evaluate the backbone structural validity of generated conformations. A conformation is consider to have *Clash* if the distance between any two α-carbon is less than two times the van der Waals radius of α-carbon with an overlap tolerance: $\delta = 2 \times 1.7 - 0.4 = 3.0$ Å. For *bond break*, we surveyed the maximum adjacent α-carbon distances in the reference MD trajectories of fast-folding proteins and BPTI. We found the maximum distance is around 4.19 Å consistent cross different protein and therefore use it as a cutoff value. Any conformation contains distance between adjacent α-carbon greater than 4.19 Å is consider having *bond break* and an invalid structure. For a rigorous evaluation, we report the overall validity VAL-CA as the fraction of conformations without *Clash* and *Bond break*.

D Additional Experimental Results

D.1 Ablation Studies on Force and Energy Guidance

In addition to different levels of *force guidance* shown in the main text, conducted similar ablation study on *energy guidance* and on other proteins in fast-folding protein datasets. As shown in Figure S2, similar to *force guidance*, *energy guidance* also improve sample energy while maintaining the level of sample diversity.

In addition, we compare the fast-folding related metrics in Table S4. Notably, validity score (VAL-CA) improves after adding energy and force guidance.

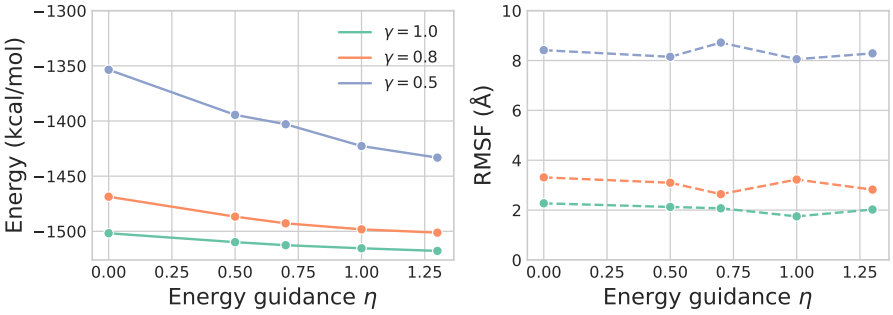

Figure S2. Energy (left) and diversity (RMSF, right) of sampled conformation for WW domain with various levels of energy guidance (η) and sequence condition (γ).

Table S4. Ablation studies on different levels of force and energy guidance on CONFDIFF, performance is evaluated on fast-folding proteins. mean/median values are reported.

	J-S DISTANCE (↓)				VAL-CA	$\text{RMSE}_{\text{CONTACT}}$	RMSF
MODELS	PWD	RG	TIC	TIC-2D	(↑)	(↓)	(Å)
			CONFDIFF-ENERGY				
$\gamma = 1.0, \eta = 0.0$	0.41/0.44	0.40/0.44	0.47/0.51	0.59/0.61	0.97/0.98	4.87/4.69	2.3/2.0
$\gamma = 1.0, \eta = 0.5$	0.39/0.39	0.37/0.38	0.45/0.51	0.58/0.62	0.97/0.99	4.75/4.27	3.3/2.6
$\gamma = 1.0, \eta = 0.7$	0.39/0.41	0.37/0.40	0.45/0.51	0.58/0.62	0.97/0.99	4.69/4.32	3.7/3.0
$\gamma = 1.0, \eta = 1.0$	0.39/0.41	0.37/0.38	0.45/0.50	0.58/0.62	0.98/0.99	4.62/4.33	4.0/3.0
$\gamma = 1.0, \eta = 1.3$	0.40/0.39	0.37/0.36	0.45/0.49	0.58/0.61	0.98/0.99	4.56/4.33	4.5/3.4
$\gamma = 0.8, \eta = 0.0$	0.33/0.33	0.31/0.31	0.45/0.46	0.61/0.63	0.92/0.93	4.09/4.10	4.2/4.0
$\gamma = 0.8, \eta = 0.5$	0.34/0.36	0.31/0.32	0.41/0.44	0.56/0.60	0.96/0.98	4.10/3.98	4.8/4.6
$\gamma = 0.8, \eta = 0.7$	0.35/0.36	0.32/0.32	0.42/0.44	0.56/0.60	0.97/0.98	4.09/4.05	5.0/4.4
$\gamma = 0.8, \eta = 1.0$	0.36/0.35	0.32/0.31	0.42/0.44	0.56/0.60	0.97/0.98	4.07/4.09	5.3/5.1
$\gamma = 0.8, \eta = 1.3$	0.36/0.37	0.33/0.34	0.42/0.43	0.56/0.60	0.97/0.98	4.04/4.08	5.8/5.2
$\gamma = 0.5, \eta = 0.0$	0.29/0.26	0.26/0.23	0.41/0.41	0.57/0.62	0.84/0.88	3.48/3.68	8.2/8.3
$\gamma = 0.5, \eta = 0.5$	0.33/0.33	0.31/0.29	0.41/0.40	0.57/0.60	0.92/0.96	3.51/3.16	8.7/8.3
$\gamma = 0.5, \eta = 0.7$	0.35/0.33	0.33/0.30	0.41/0.40	0.57/0.60	0.94/0.97	3.60/3.26	8.6/8.7
$\gamma = 0.5, \eta = 1.0$	0.36/0.34	0.35/0.31	0.41/0.39	0.57/0.60	0.95/0.96	3.68/3.42	9.0/9.2
$\gamma = 0.5, \eta = 1.3$	0.37/0.35	0.36/0.33	0.41/0.40	0.58/0.60	0.95/0.96	3.76/3.51	8.9/9.1
			CONFDIFF-FORCE				
$\gamma = 1.0, \eta = 0.0$	0.41/0.44	0.40/0.44	0.47/0.51	0.59/0.61	0.97/0.98	4.87/4.69	2.3/2.0
$\gamma = 1.0, \eta = 0.5$	0.38/0.42	0.36/0.41	0.46/0.50	0.58/0.64	0.97/0.98	4.66/4.27	3.3/3.1
$\gamma = 1.0, \eta = 0.7$	0.38/0.42	0.36/0.40	0.46/0.50	0.58/0.63	0.97/0.98	4.58/4.22	3.3/3.0
$\gamma = 1.0, \eta = 1.0$	0.38/0.42	0.36/0.40	0.46/0.51	0.59/0.64	0.98/0.99	4.46/4.11	3.3/2.8
$\gamma = 1.0, \eta = 1.3$	0.38/0.41	0.36/0.40	0.47/0.52	0.59/0.64	0.98/0.99	4.37/3.97	3.3/2.8
$\gamma = 0.8, \eta = 0.0$	0.33/0.33	0.31/0.31	0.45/0.46	0.61/0.63	0.92/0.93	4.09/4.10	4.2/4.0
$\gamma = 0.8, \eta = 0.5$	0.32/0.30	0.30/0.27	0.42/0.46	0.56/0.59	0.95/0.96	3.87/3.64	4.6/4.3
$\gamma = 0.8, \eta = 0.7$	0.32/0.29	0.30/0.26	0.42/0.45	0.57/0.60	0.96/0.97	3.82/3.60	4.5/4.2
$\gamma = 0.8, \eta = 1.0$	0.32/0.29	0.29/0.26	0.43/0.46	0.58/0.61	0.97/0.98	3.71/3.51	4.6/4.4
$\gamma = 0.8, \eta = 1.3$	0.32/0.30	0.30/0.28	0.43/0.46	0.58/0.60	0.98/0.98	3.67/3.58	4.7/4.1
$\gamma = 0.5, \eta = 0.0$	0.29/0.26	0.26/0.23	0.41/0.41	0.57/0.62	0.84/0.88	3.48/3.68	8.2/8.3
$\gamma = 0.5, \eta = 0.5$	0.31/0.28	0.28/0.25	0.41/0.41	0.58/0.62	0.91/0.95	3.38/3.32	8.5/8.6
$\gamma = 0.5, \eta = 0.7$	0.31/0.28	0.28/0.26	0.41/0.41	0.58/0.62	0.93/0.97	3.32/3.24	8.3/8.7
$\gamma = 0.5, \eta = 1.0$	0.31/0.29	0.29/0.26	0.42/0.41	0.58/0.63	0.96/0.97	3.45/3.27	8.5/9.1
$\gamma = 0.5, \eta = 1.3$	0.32/0.30	0.30/0.28	0.41/0.40	0.58/0.63	0.97/0.98	3.52/3.29	8.4/8.5

D.2 Additional Discussion on Fast-folding Results

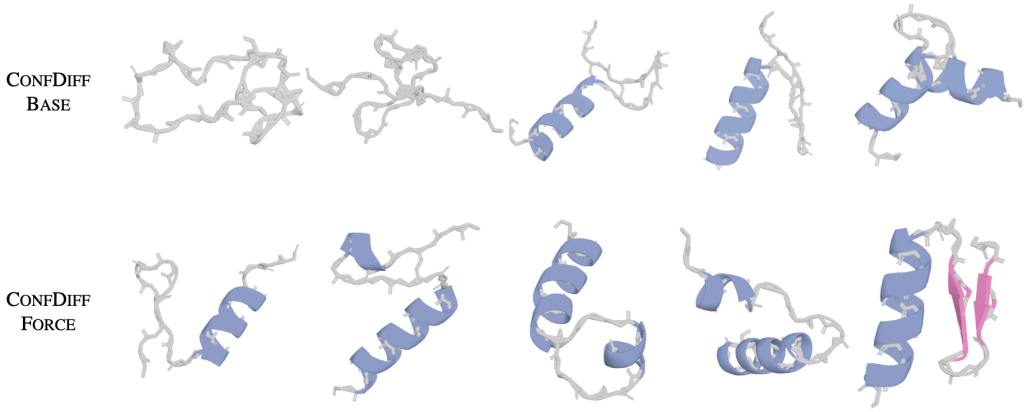

Figure S3. Five worst structures with highest residue clash rate. Both the cartoon representation and backbone bonds are shown to better depict the clashes.

We find that Jensen-Shannon (JS) metrics alone are insufficient to reflect the quality and diversity of sampled conformations. Note that the JS distance here is calculated and averaged over each dimension (e.g., α-carbons), which is not able to capture the bond breaking and steric clash. With the introduction of energy/force guidance, irrational conformations are rectified and structures tend to contract towards the local minimum. This is the main factor leading to a performance drop in distributional metrics.

Empirically, CONFDIFF-base sometimes samples poor structures with steric clashes, while energy/force guidance strategies enable sampling high-fidelity conformations (i.e., high VAL-CA scores) with slightly worse JS metrics. As shown in Figure 2 in our paper, RMSF (reflecting structural diversity) values fluctuate with various force guidance strengths, while the potential energy keeps improving. This also explains the fluctuation of JS metrics with improved VAL-CA scores after energy/force guidance.

Here we show some examples to compare sampled structures from CONFDIFF-base and CONFDIFF-force. Among 1000 samples from CONFDIFF-base ($\gamma = 0.5$) and CONFDIFF-force ($\gamma = 0.5$, $\eta = 1.3$), we select five worst conformations (with highest clash rate) shown in the Figure S3. CONFDIFF-force samples exhibit fewer steric clashes, demonstrating more realistic protein structures, despite slightly worse JS distance. Therefore, it is more appropriate to take into account all metrics for a comprehensive assessment of the generated conformation ensemble.

D.3 Sample Distribution over TIC Components for All Fast-Folding Proteins

Here we include sample distributions over the first two TIC components for all twelve fast-folding proteins.

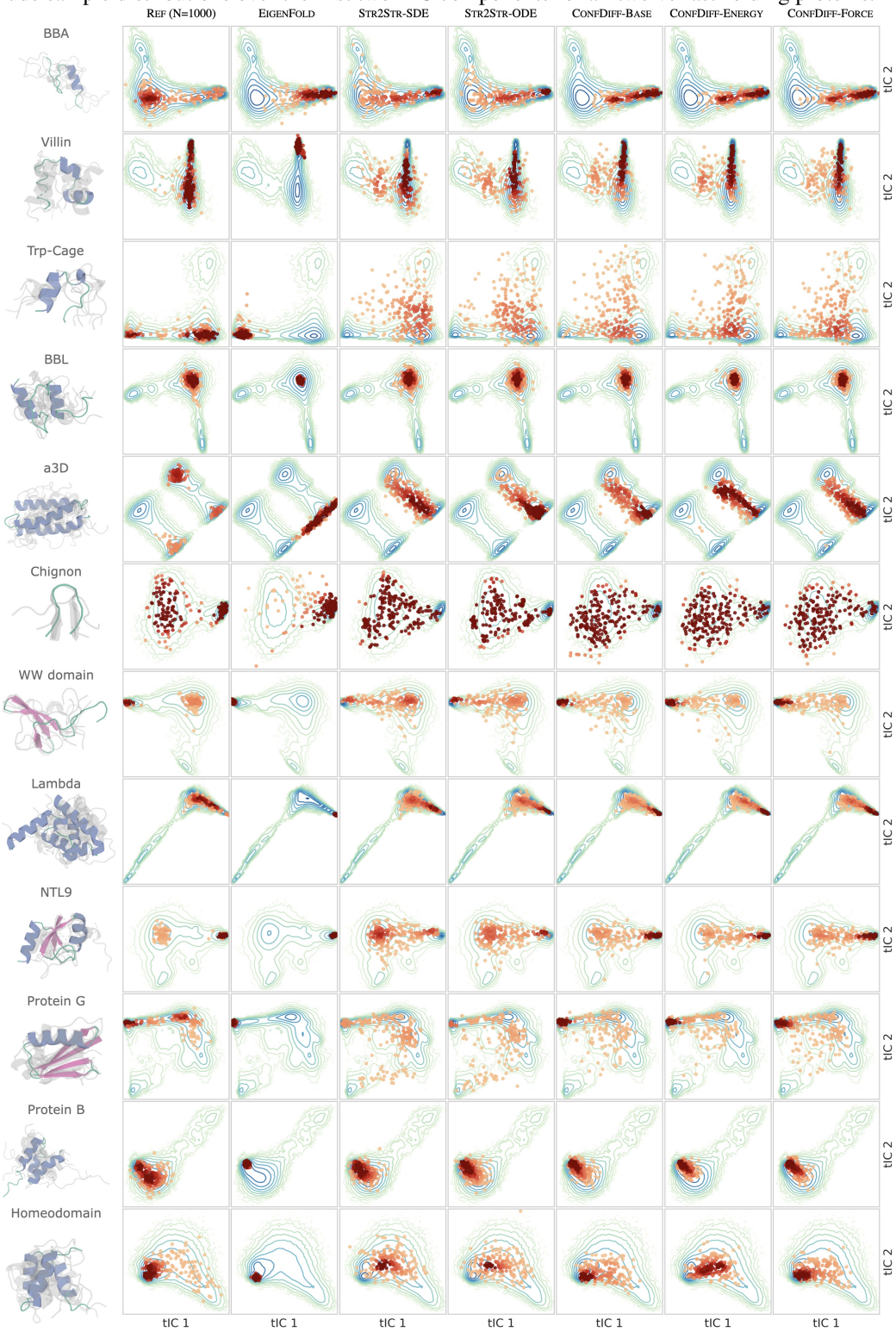

Figure S4. Sample distribution over the first two TIC components for twelve fast-folding proteins. Contours indicate the full reference MD sample density at log-scale.

D.4 Additional Results for BPTI

We list the best RMSD scores to each five cluster centers in Table D.4. CONFDIFF-FORCE and EIGENFOLD can most accurately predict the structures of metastable states. Especially, CONFDIFF-FORCE out-performs all other models in predicting the remote cluster 3.

Table S5. Detailed best RMSD to each cluster center. Unit: Å.

MODEL	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5
REF (N=1000)	0.79	0.79	2.19	1.02	0.74
EIGENFOLD	**0.86**	1.69	2.45	**1.34**	**0.99**
STR2STR-ODE	1.85	1.94	2.46	2.13	1.9
STR2STR-SDE	2.04	2.01	2.80	2.34	2.12
CONFDIFF-BASE	1.02	1.72	2.38	1.5	1.12
CONFDIFF-ENERGY	1.16	1.72	2.38	1.49	1.11
CONFDIFF-FORCE	**0.86**	**1.58**	**2.12**	1.44	1.03

Similar to fast-folding, we visualize the sample distribution over the first two TIC components (Figure S5). Despite that EIGENFOLD predicts structures for cluster 2, 3, 4 better STR2STR and two CONFDIFF models, its TICA plot shows it is concentrated around cluster 1 while other models have better sample coverage.

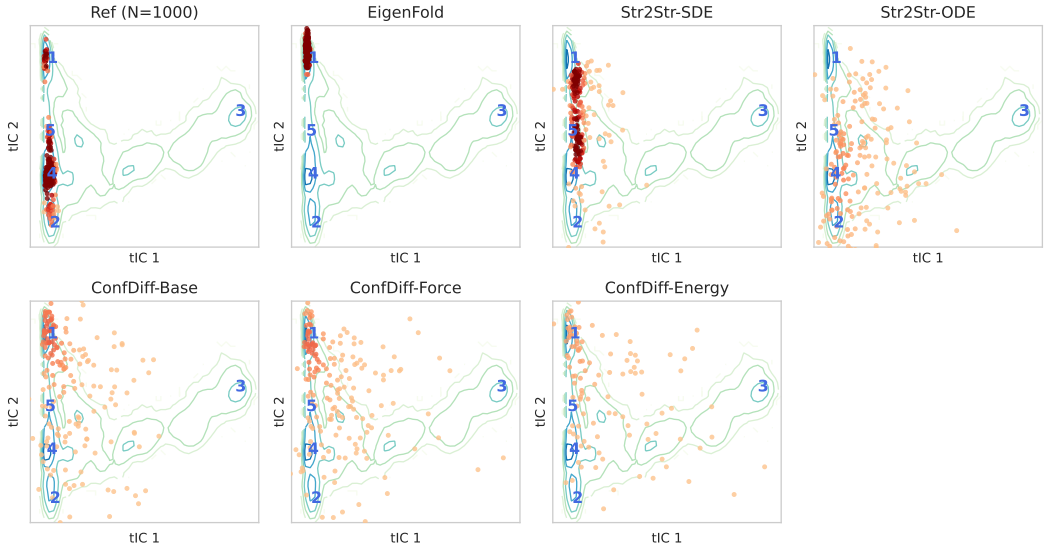

Figure S5. Model generated BPTI conformations projected to the first two TIC components. Contour lines represent the ground truth probability density at the log scale.

D.5 CONFDIFF-BASE Results on Apo-holo Dataset

Apo-holo is a dataset consists of 90 proteins with known conformation changes upon ligand-binding and each has two conformations (*apo*, the unbound state; *holo*, the bound state) determined by experiments. This dataset introduces a challenging task of predicting distinct conformations of proteins with various sizes and structural changes (Jing et al., 2023; Saldaño et al., 2022). Due to the high computation cost of evaluating the energy and force for mid-to-large size structure using `OpenMM`, we only compare the performance of our *classifier-free* model with baselines.

To evaluate the performance on predicting two distinct conformations, we calculated TM-ensemble score (TMENS) as in (Jing et al., 2023; Lu et al., 2024):

$$\text{TMENS} = 1/2 \times \left(\max_i \left\{ \text{TMscore}(\text{sample } i, \textit{apo}) \right\} + \max_i \left\{ \text{TMscore}(\text{sample } i, \textit{holo}) \right\} \right)$$

where TMscore is a normalized structural similarity score between 0 and 1 (Zhang & Skolnick, 2004). Results are summarized in Table S6 and Figure S6. None of the model shows superb performance as most of the samples perform worse than a perfect folding model, indicating remaining challenges for current conformation models to sample distinct structures with potentially large conformation changes. Among models compared, EIGENFOLD and CONFDIFF-BASE ($\gamma = 1.0$) show better performance that likely due to better prediction of a "folded" state for a given protein; while STR2STR models show lower performance potentially due to the lack of sequence information during perturbation for accurate folding prediction.

Table S6. Apo-holo results (90 cases). TM-DIV is the mean pairwise TMscore between samples.

	TMENS	RMSF	TM-DIV
EIGENFOLD	0.846/0.875	3.44/1.19	0.903/0.958
STR2STR-ODE	0.775/0.790	1.92/1.94	0.894/0.907
STR2STR-SDE	0.668/0.684	4.81/4.70	0.677/0.663
CONFDIFF-BASE ($\gamma = 1.0$)	0.843/0.874	3.35/2.27	0.847/0.883
CONFDIFF-BASE ($\gamma = 0.9$)	0.834/0.869	4.12/2.91	0.790/0.813
CONFDIFF-BASE ($\gamma = 0.8$)	0.818/0.843	5.55/4.60	0.693/0.703

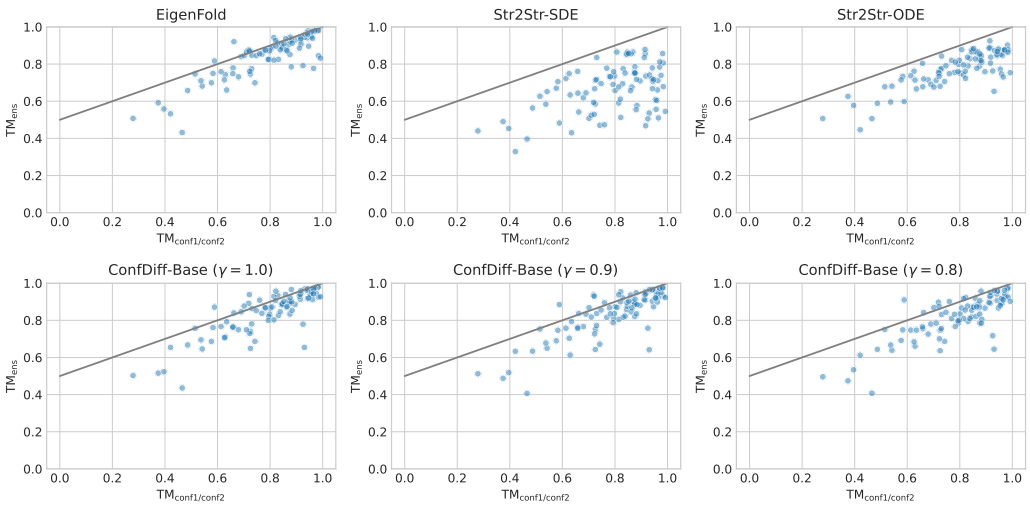

Figure S6. TMens vs TMscore between *apo* and *holo* structures. A small $t = 0.1$ selected for STR2STR models. The slide line indicators the baseline performance of a perfect folding model.

D.6 CONFDIFF Results on Sars-CoV-2

In order to fully understand the method's generalization capabilities, we did some proof-of-concept experiments to extend proposed method to challenging cases of large proteins - the SARS-CoV-2 case studies introduced in Distributional Graphormers (DiG) (Zheng et al., 2023). This dataset contains two important proteins in drug discovery: the spike receptor-binding proteins (RBD, 194 amino acids) and the main protease (Mpro, 306 amino acids). MD simulations for both proteins are publicly available and are used as reference during evaluation.

Preliminary results showed that protein sequence information using representations from ESMFold (Lin et al., 2022) is not able to predict reasonable structures of RBD and Mpro, while AlphaFold predictions are close to the experimental structures in Table S7.

Table S7. The RMSD and TMscore of RBD and Mpro from different sequence representations.

	RMSD	TMSCORE
ESMFOLD_RBD	20.55	0.18
ESMFOLD_MPRO	13.96	0.59
ALPHAFOLD_RBD	2.55	0.91
ALPHAFOLD_MPRO	3.28	0.93

For quantitative comparison, we calculated the following metrics: 1) average RMSD to experimental structures (PDB IDs are RBD: 6LU7, Mpro: 6M0J); 2) the JS-TIC2D to reference distribution; 3) performances on the binary classification of TICA states similar to DiG: the joint space is discretized into 50×50 states and those with/without ground truth conformations are positive/nagtive states; 4) diversity as pairwise RMSD between samples (RMSF) We summarized the metrics below in Table S8, with some major findings:

Table S8. Results on Sars-CoV-2. The results for classifier-free, energy guidance, and force guidance are labeled as CONFDIFF-BASE, CONFDIFF-FORCE, and CONFDIFF-ENERGY. All values are shown as mean.

MODEL	RMSD (Å)	JS-TIC2D ↓	BINARY PREDICTION OF TIC2D STATES			RMSF (Å)
			PRECISION ↑	RECALL ↑	F1 ↑	
RBD						
CONFDIFF-BASE	1.396	**0.822**	0.72	0.147	0.244	1.55
CONFDIFF-ENERGY	1.341	0.826	0.647	0.133	0.221	1.564
CONFDIFF-FORCE	1.385	0.825	**0.765**	**0.154**	**0.257**	1.606
MPRO						
CONFDIFF-BASE	1.475	0.785	1	0.125	0.221	1.692
CONFDIFF-ENERGY	1.448	0.794	1	0.116	0.208	1.722
CONFDIFF-FORCE	1.559	**0.768**	1	**0.132**	**0.234**	1.828

- All the models can accurately predict the structure for both RBD and Mpro with mean RMSD<1.6 Å. The samples are scattered near the major peaks in the reference distribution as a single cluster. For RBD, the samples spread out in the same direction as the reference distributions.
- Adding energy or force guidance does not significantly impact the metrics. While CONFDIFF-Force showed some positive signs on slightly improved metrics (e.g., JS-TIC2D for Mpro and Recall & F1 scores for both proteins).
- The current prototype is only fine-tuned on the AF2 representation with two proteins (RBD and Mpro). Lack of training on large datasets could significantly impact model's performance. We could train the base model on the entire PDB structures with AF2 representations in the future and we believe the model's performance would greatly improve.
- Nevertheless, we aim to use these experiments to show that 1) proposed methods can scale up and apply to large proteins with real-world applications; 2) even with much simplification on experiments, the results showed some promising signs on proposed results.