Progressive Inference-Time Annealing of Diffusion Models for Sampling from Boltzmann Densities

Anonymous Authors[1]

Abstract

Sampling efficiently from a target unnormalized probability density remains a core challenge, with relevance across countless high-impact scientific applications. A promising approach towards this challenge is the design of amortized samplers that borrow key ideas, such as probability path design, from state-of-the-art generative diffusion models. However, all existing diffusion-based samplers remain unable to draw samples from distributions at the scale of even simple molecular systems. In this paper, we propose PROGRESSIVE INFERENCE-TIME ANNEALING (PITA) a novel framework to learn diffusion-based samplers that combines two complementary interpolation techniques: I.) Annealing of the Boltzmann distribution and II.) Diffusion smoothing. PITA trains a sequence of diffusion models from high to low temperatures by sequentially training each model at progressively higher temperatures, leveraging engineered easy access to samples of the temperature-annealed target density. In the subsequent step, PITA enables simulating the trained diffusion model to *procure training samples at a lower temperature* for the next diffusion model through inference-time annealing using a novel Feynman-Kac PDE combined with Sequential Monte Carlo. Empirically, PITA enables, for the first time, equilibrium sampling of N-body particle systems, Alanine Dipeptide, and tripeptides in Cartesian coordinates with dramatically lower energy function evaluations.

1. Introduction

The problem of sampling from an unnormalized target probability distribution arises in numerous areas of natural sciences, including computational biology, chemistry, physics, and materials science (Frenkel & Smit, 2023; Liu, 2001; Ohno et al., 2018; Stoltz et al., 2010). In many of these high-impact scientific settings, the problem's complexity stems from operating in molecular systems where the unnormalized target (Boltzmann) distribution at a low temperature of interest is governed by a highly non-convex and non-smooth energy function, under which there is limited to no available data (Hénin et al., 2022). As a result, the challenging sampling problem necessitates solving an equally hard exploration problem: finding the modes—in correct proportion—of the target distribution.

To address the general sampling problem, extensive research has been dedicated to Markov chain Monte Carlo methods (MCMC), Sequential Monte Carlo (SMC), and, particularly in physical systems, Molecular Dynamics (MD) (Leimkuhler & Matthews, 2015). To enhance scalability, Monte Carlo approaches often employ an interpolating sequence of probability distributions that transitions from an easily sampled reference distribution to the desired target distribution via annealing/tempering strategies. This powerful concept underlies methods such as parallel tempering (Swendsen & Wang, 1986), Annealed Importance Sampling (Jarzynski, 1997; Neal, 2001), and SMC samplers (Del Moral et al., 2006). MD, conversely, involves integrating equations of motion using finely discretized time steps. Despite their effectiveness, both classes of methods possess inherent limitations that complicate their application to practical systems of interest: annealing modifies the masses of distribution modes depending on their widths (a phenomenon often referred to as mass teleportation, (Woodard et al., 2009)), while MD requires computationally expensive time discretization on the order of femtoseconds to simulate millisecond-scale phenomena.

Diffusion-based samplers are an alternative and emergent class of sampling techniques (Zhang & Chen, 2022; Vargas et al., 2023; Akhound-Sadegh et al., 2024; Berner et al., 2022; Blessing et al., 2024; Havens et al., 2025) exploiting modern developments in generative modeling. They sample complex multi-modal distributions by leveraging a prescribed interpolating probability path. However, instead of relying on annealing, these samplers utilize a noising mechanism which theoretically enjoys favorable mixing

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

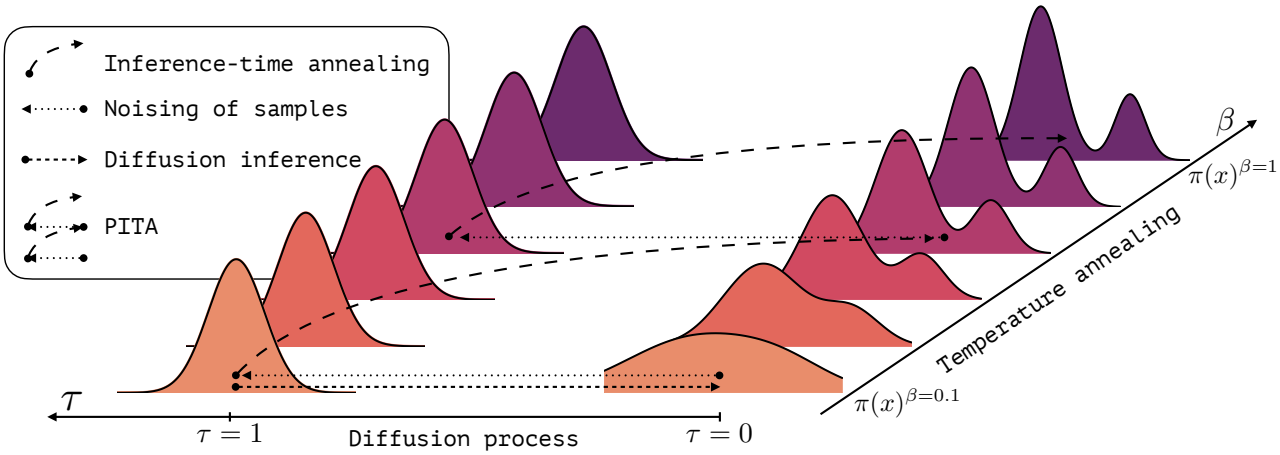


Figure 1: Illustration of the proposed PITA framework combining two complementary processes: temperature annealing of the target Boltzmann density and the diffusion process applied to the collected samples. Annealed inference allows for decreasing the temperature (increasing β) of a trained diffusion model, thus generating samples from the annealed target. These samples can be reused for training a lower-temperature diffusion model.

properties compared to annealing (Chen et al., 2023).

Diffusion-based samplers, despite their appeal, have not yet proven effective for even small molecular systems in Cartesian coordinates. This is primarily because of the absence of training data to accurately approximate the logarithmic gradients of the marginal densities, i.e. the Stein scores – a challenge distinct from generative modeling settings. Additionally, standard training objectives, such as reverse Kullback–Leibler, are mode dropping and often yield too high-variance score estimates for stable training (Blessing et al., 2024). Crucially, current diffusion-based samplers require too many energy function evaluations for training. Indeed, when normalized by the number of energy evaluations, carefully tuned MCMC methods with parallel tempering are empirically competitive with, if not superior to, state-of-the-art diffusion-based samplers (He et al., 2025).

Main Contributions. In this paper, we introduce PROGRESSIVE INFERENCE-TIME ANNEALING (PITA), a novel framework for training diffusion models to sample from Boltzmann distributions. PITA leverages two complementary interpolation techniques to significantly enhance training scalability: temperature annealing (increasing the system's temperature) and interpolation along a conventional diffusion model's probability path. This combination is motivated by a learning framework designed to exploit their distinct advantages: temperature annealing mixes modes by lowering high energy barriers, while diffusion paths avoid mass teleportation.

Annealing the target distribution transforms the challenging sampling problem into an easier one by removing high-energy barriers and flattening it. This crucial step enables cheap collection of an initial high-temperature dataset via classical MCMC, which in turn facilitates the efficient training of an initial diffusion model. Subsequently, we define a novel Feynman–Kac PDE that, when combined with SMC-based resampling, allows us to simulate the trained diffusion model's inference process to produce asymptotically unbiased samples at a lower temperature. This effectively allows us to train the next diffusion model, enabling the progressive and stable training of a sequence of diffusion models until the target distribution is reached, as illustrated in Fig. 1.

We test the empirical performance of PITA on standard N-body particle systems and short peptides in Alanine Dipeptide and tripeptides. Empirically PITA not only achieves state-of-the-art performance in all these benchmarks but is the first diffusion-based sampler that scales to our considered peptides in their native Cartesian coordinates. More importantly, we demonstrate progressing down our designed ladder of diffusion models leads to significantly lower energy evaluations compared to MD, which is a step towards realizing the promise of amortized samplers for accelerating equilibrium sampling.

2. Background

2.1. Diffusion models

A diffusion process defines an interpolating path between an easy-to-sample reference density, such as a multi-variate Normal, and a desired target distribution $\pi(x)$. When samples from the target distribution are available, it is possible to generate samples from intermediate marginals of the diffusion process $p_\tau(x)$ through the following Gaussian convolution:

$$p_\tau(x) = \big(\pi * \mathcal{N}(\alpha_\tau y, \sigma_\tau^2 \mathbb{1})\big)(x) = \int dy\, \mathcal{N}(x \,|\, \alpha_\tau y, \sigma_\tau^2 \mathbb{1})\pi(y). \tag{1}$$

As a result, this means that samples from $p_\tau(x)$ can be generated as $x_\tau = \alpha_\tau y + \sigma_\tau \varepsilon$, where $y \sim \pi(y)$ and $\varepsilon \sim \mathcal{N}(0, \mathbb{1})$. Selecting specific schedules for α_τ and σ_τ one can ensure the following boundary conditions. For $\tau = 0$, $\alpha_\tau = 1, \sigma_\tau = 0$ and $p_{\tau=0}(x) = \pi(x)$, i.e. the marginal matches the target distribution. For $\tau = 1$, $\alpha_\tau = 0, \sigma_\tau = 1$ and $p_{\tau=1}(x) = \mathcal{N}(0, \mathbb{1})$, i.e. the marginal matches the standard multivariate normal distribution.

Importantly, despite the simplicity of sample generation, the evaluation of density $p_\tau(x)$ is not straightforward, and one has to use deep learning models to approximate either scores $\nabla \log p_\tau(x)$ or marginal densities $p_\tau(x)$. Furthermore, the model density $p_\tau(x)$ or scores $\nabla \log p_\tau(x)$ can be used to generate new samples from $\pi(x)$ using the reverse-time SDE. In particular, the marginals introduced in Eq. (1) describe the marginal densities of the forward-time (Ornstein–Uhlenbeck) SDE

$$dx_\tau = \underbrace{\frac{\partial \log \alpha_\tau}{\partial \tau} x_\tau}_{:=a_{1-\tau} x_\tau} d\tau + \underbrace{\sqrt{2\sigma_\tau^2 \frac{\partial}{\partial \tau} \log \frac{\sigma_\tau}{\alpha_\tau}}}_{:=\zeta_{1-\tau}} d\overline{W}_\tau \,, \quad x_{\tau=0} \sim \pi(x) \,, \tag{2}$$

where $\overline{W}_\tau$ is the standard Wiener process, and marginals follow the Fokker-Planck PDE. After inverting the time variable in this PDE, i.e. $t = 1 - \tau$, the time-evolution of marginals $p_t(x)$ is

$$\begin{aligned} \frac{\partial p_t(x)}{\partial t} &= \langle \nabla, p_t(x)(a_t x) \rangle - \frac{\zeta_t^2}{2} \Delta p_t(x) \\ &= -\left\langle \nabla, p_t(x)(-a_t x + \frac{\zeta_t^2}{2} \nabla \log p_t(x)) \right\rangle , \end{aligned} \tag{3}$$

which shows that one can sample from the marginals $\{p_t(x)\}_{t\in[0,1]}$ via simulating the following SDE

$$dx_t = \left(-a_t x + \frac{\zeta_t^2}{2}(1 + \xi_t) \nabla \log p_t(x_t) \right) dt + \zeta_t \sqrt{\xi_t} dW_t \,. \tag{4}$$

While the marginals are correct for any $\xi_t > 0$, there are two important special cases: for $\xi_t \equiv 1$, the equation becomes the reverse-SDE with the same path-measure as Eq. (2), and, for $\xi_t \equiv 0$, the SDE becomes an ODE. In practice, this SDE is simulated using either the density model $\exp(-U_t(x;\eta)) \propto p_t(x)$ (Du et al., 2023) or the score model $s_t(x;\theta) \approx \nabla \log p_t(x)$ (Song et al., 2021).

2.2. Annealing

Annealing defines a family of "simpler" problems when we have access to the unnormalized density by interpolating or scaling the target log-density (negative energy). Formally speaking, given the unnormalized density $\pi(x)$, the annealed density is defined as

$$\pi^\beta(x) := \frac{\pi(x)^\beta}{Z_\beta} \,, \quad Z_\beta = \int dx \, \pi(x)^\beta \,, \tag{5}$$

where β is the inverse temperature parameter (i.e., $\beta = 1/T$) controlling the smoothness of the target density. Thus, for high temperature $T > 1$ (hence, $\beta < 1$), the target density becomes smoother and easier to explore via MCMC algorithms. Importantly, getting the unnormalized density for $\pi^\beta(x)$ can be simply achieved by raising the unnormalized density $\pi(x)$ to the power β.

2.3. Feynman-Kac Formula

The Feynman-Kac Partial Differential Equation (PDE) is a generalization of the Fokker-Planck PDE and is defined as follows

$$\begin{aligned} \frac{\partial p_t(x)}{\partial t} &= -\langle \nabla, p_t(x) v_t(x) \rangle + \frac{\zeta_t^2}{2} \Delta p_t(x) \\ &\quad + p_t(x) \big(g_t(x) - \mathbb{E}_{p_t(x)} g_t(x) \big) \,, \end{aligned} \tag{6}$$

where the first term corresponds to the probability mass transport along the vector field $v_t(x)$, the second term corresponds to the stochastic moves of samples according to the Wiener process W_t, and the last term is responsible for reweighting the samples according to a coordinate dependent weighting function $g_t(x)$. For any test-function $\varphi(x)$, the Feynman-Kac formula relates its expected value to the expectation over the SDE trajectories x_t, i.e.

$$\begin{aligned} \mathbb{E}_{p_T(x)}[\varphi(x)] &= \frac{1}{Z_T} \mathbb{E}\Big[e^{\int_0^T dt \; g_t(x_t)} \varphi(x_T) \Big] \,, \\ \text{where } dx_t &= v_t(x_t) dt + \zeta_t dW_t \,, \; x_0 \sim p_0(x) \,, \end{aligned} \tag{7}$$

and Z_T is a normalization constant independent of x. In practice, the exponential term is computed as a "weight" w_t of the corresponding sample x_t and can be integrated in parallel with the simulation,

$$\begin{aligned} dx_t &= v_t(x_t) dt + \zeta_t dW_t \,, \quad d \log w_t = g_t(x_t) dt \,, \\ &\text{initialized as } \; x_0 \sim p_0(x) \,, \quad \log w_0 = 0 \,. \end{aligned} \tag{8}$$

Finally, one can estimate the normalization constant Z_T by considering $\varphi(x) \equiv 1$ in Eq. (7) and get the biased but consistent Self-Normalized Importance Sampling (SNIS) estimator (Liu, 2001), i.e.

$$\begin{aligned} \frac{1}{Z_T} \mathbb{E}\Big[e^{\int_0^T dt \; g_t(x_t)} \varphi(x_T) \Big] &= \frac{\mathbb{E} e^{\int_0^T dt \; g_t(x_t)} \varphi(x_T)}{\mathbb{E} e^{\int_0^T dt \; g_t(x_t)}} \\ &\approx \sum_{i=1}^n \frac{w_T^i}{\sum_{j=1}^n w_T^j} \varphi(x_T^i) \,, \end{aligned} \tag{9}$$

where (x_T^i, w_T^i) are the samples from the simulation of the SDE in Eq. (8).

3. Progressive Inference-Time Annealing

In this section, we combine diffusion and annealing processes into an efficient learning algorithm for sampling from the target density $\pi(x)$. To design this method, we build on the fact that diffusion and annealing are complementary ways to simplify or "smoothen" the target distribution (see Fig. 1). Namely, for the high-temperature version of the target distribution $\pi^{\beta_i}(x)$, we assume having samples from $\pi^{\beta_i}(x)$ and learn the density model of the marginals defined by the diffusion process (see Section 3.2). For instance, this can be done by running MCMC chains that face little challenge mixing in high temperatures. For the given density model of the diffusion process, we perform annealing of all the marginals and generate samples from the lower temperature target $\pi^{\beta_{i+1}}(x)$, $\beta_{i+1} < \beta_i$ (see Section 3.1). We detail every step of our method in the following subsections.

3.1. Inference-Time Annealing

In this section, we discuss the inference-time annealing process, which allows us to modify the trained diffusion model

to generate samples from the lower temperature target density. Namely, for a diffusion process with marginals $p_t(x)$ and the end-point $p_{t=1}(x) = \pi^{\beta_i}(x)$, we assume having two models: a score model $s_t(x;\theta) \approx \nabla \log p_t(x)$ and an energy-based model $U_t(x;\eta) \approx -\log p_t(x) + \text{const}$ with parameters θ and η respectively. Given the score and the energy models trained to sample from a higher temperature density $\pi^{\beta_i}(x)$, we define a new sequence of marginals that correspond to the Boltzmann density of the energy model but with a lower temperature

$$q_t(x) \propto \exp(-\beta_{i+1}/\beta_i U_t(x;\eta)) \,,$$
$$q_{t=1}(x) \propto \exp(-U_{t=1}(x;\eta))^{\beta_{i+1}/\beta_i} \approx (\pi(x)^{\beta_i})^{\beta_{i+1}/\beta_i} \,. \tag{10}$$

The following proposition derives the Feynman-Kac PDE that describes the time-evolution of the marginals $q_t(x)$ and allows for importance sampling via the Feynman-Kac formula.

Proposition 1. [Inference-time Annealing] *Annealed density of the energy-based model $q_t(x) \propto \exp(-\gamma U_t(x;\eta))$ matches the marginal densities of the following SDE*

$$dx_t = \left(-a_t x_t + \frac{\zeta_t^2}{2}(s_t(x_t) - \gamma \xi_t \nabla U_t(x_t;\eta))\right)dt + \zeta_t \sqrt{\xi_t} dW_t \,, \quad x_0 \sim q_{t=0}(x) \tag{11}$$

$$d\log w_t = \Bigg[\frac{\zeta_t^2}{2}\langle \nabla, s_t(x_t)\rangle - \gamma \left\langle \nabla U_t(x_t;\eta), -a_t x_t + \frac{\zeta_t^2}{2} s_t(x_t)\right\rangle - \gamma \frac{\partial U_t(x_t;\eta)}{\partial t}\Bigg] dt \,, \tag{12}$$

where $s_t(x)$ is any vector field, a_t, ζ_t, ξ_t are analogous to parameters from Eq. (4), and the sample weights w_t correspond to the SNIS estimator of the Feynman-Kac formula in Eq. (9).

See Appendix A.1 for the proof. Intuitively, this result defines an importance sampling scheme, where Eq. (11) generates samples from the proposal distribution and Eq. (12) integrates the density ratio between the sampled proposal and the target density $q_t(x)$. Different choices of the vector field $s_t(x)$ and the noise schedule ξ_t yield different proposal distributions. Theoretically, one can choose different parameters a_t, ζ_t as well, but below we argue for setting them according to Eq. (2).

The dynamics in Proposition 1 is not unique, and there exists a continuous family of PDEs that follow the marginals $q_t(x) \propto \exp(-\gamma U_t(x;\eta))$. We motivate this dynamics as minimizing the variance of the weights for the case when there is no annealing ($\gamma = 1$). Indeed, if the trained EBMs and score models approximate the diffusion process perfectly, then, for $\gamma = 1$, the weights become constant, so SNIS equally weights all the samples; thus, eliminating the need to resample at all. We formalize this result in the following proposition (see Appendix A.1 for the proof).

Proposition 2. [Convergence to Diffusion] *For $\gamma = 1$ and perfect models $s_t(x) = -\nabla U_t(x;\eta) = \nabla \log p_t(x)$, the variance of the weights in Proposition 1 becomes zero.*

In the case of unbounded support of the target distribution, e.g. $\texttt{supp}(\pi(x)) = \mathbb{R}^d$, increasing the temperature might cause numerical instabilities. Indeed, $\pi(x)^{\beta=0} \propto \texttt{Uniform}(\mathbb{R}^d)$ is not normalizable. To avoid this potential issue, in Appendix A.2, we consider geometric averaging between some simple prior and the target densities, e.g. $\mathcal{N}(0, \mathbb{1})^{(1-\beta)}\pi(x)^{\beta}$.

Integrating the dynamics from Propositions 1 and 3 we generate a set of weighted samples $\{(x_{t=1}^k, w_{t=1}^k)\}_{k=1}^K$ that converge to the samples from $q_{t=1}(x)$ when $K \to \infty$. In practice, this density is defined as the Boltzmann distribution of the corresponding energy model, i.e. $q_{t=1}(x) \propto \exp(-\beta_{i+1}/\beta_i \cdot U_{t=1}(x;\eta))$, which approximates $\pi^{\beta_{i+1}}(x)$, but does not necessarily match it exactly. We discuss several possible ways to bridge this gap between the density model and the target density in Appendix B.

3.2. Training using PITA

The proposed algorithm consists of interleaving the inference-time annealing (described in the previous two subsections) and model training on the newly generated data from the annealed target distribution, which we describe here. Throughout this stage we assume availability of samples from $\pi^{\beta_{i+1}}$, which were previously generated at the sampling stage[1].

For the target distribution $\pi^{\beta_{i+1}}(x)$, we define the diffusion process with the marginals $p_t(x)$ obtained as a convolution of the samples from the target $x \sim \pi^{\beta_{i+1}}(x)$ with the Gaussian $\mathcal{N}(\alpha_\tau x, \sigma_\tau^2 \mathbb{1})$. To learn the score function $s_t(x;\theta) \approx \nabla \log p_t(x)$, we follow the standard practice and parameterize the denoising model $D_t(x_t;\theta) = \sigma_t^2 s_t(x_t;\theta) + x_t$, which we learn via the denoising score matching (DSM) objective (Ho et al., 2020), i.e.

$$\text{Denoising Score Matching}(\theta) = \mathbb{E}_{t,x_t,x}\lambda(t)\|x - D_t(x_t;\theta)\|^2 \,, \tag{13}$$

where the expectation is taken w.r.t. samples from the annealed target $x \sim \pi^{\beta+\Delta\beta}(x)$, noised samples $x_t \sim \mathcal{N}(x_t \,|\, \alpha_{1-t}x, \sigma_{1-t}^2)$, and time parameter sampled with $\log(1-t) \sim \mathcal{N}(P_{mean}, P_{std})$ largely following Karras et al. (2022).

However, the DSM objective is not sufficient for training a good score model close to the target distribution ($\tau = 1 - t = 0$) due to the high variance of the estimator. Indeed, for $\tau = 1 - t = 0$, it has no information about the target distribution. Target Score Matching (De Bortoli et al., 2024) overcomes this issue by explicitly incorporating the score of the target unnormalized density into the objective, which is as follows

$$\text{Target Score Matching}(\theta) = \mathbb{E}_{t,x_t,x}\left[\left\|\sigma_t^2 \nabla_x \log \pi(x) + x - D_t(x_t;\theta)\right\|^2 \cdot \mathbb{I}(t \geq t_{\text{thresh}})\right] \tag{14}$$

where the expectation is taken w.r.t. the same random variables as in Eq. (13), but the time variable is restricted to the interval $[t_{\text{thresh}}, 1]$ because the variance of the objective estimator grows with the noise scale (De Bortoli et al., 2024). Notably, for larger noise scales, the Denoising Score Matching objective results in a stable training dynamics; thus, these objectives complement each other and result in a stable training dynamics across the entire time interval.

[1]For the very first iteration of our algorithm, we assume that there exist such β that samples from π^{β} can be simply collected by conventional Monte Carlo algorithms.

Algorithm 1 Training for single temperature $1/\beta_{i+1}$

Require: samples x from $\pi^{\beta_{i+1}}$.
  for training iterations **do**
    sample $\ln(\sigma_{1-t}) \sim \mathcal{N}(P_{mean}, P_{std}^2)$
    add noise $x_t \sim \mathcal{N}(x_t \mid \alpha_{1-t}x, \sigma_{1-t}^2 \mathbb{1})$
    Denoising Score Matching$(\theta) = \nabla_\theta \, \mathbb{E}_{t,x_t,x} \lambda(t) \|x - D_t(x_t;\theta)\|^2$
    Target Score Matching$(\theta) = \nabla_\theta \, \mathbb{E}_{t,x_t,x} \left[\left\| \sigma_t^2 \nabla_x \log \pi(x) + x - D_t(x_t;\theta) \right\|^2 \cdot \mathbb{I}(t \geq t_{\text{thresh}}) \right]$
    EBM Distillation$(\eta) = \nabla_\eta \, \mathbb{E}_{t,x_t,x} \lambda(t) \left\| \sigma_t^2 (-\nabla_{x_t} U_t(x_t;\eta)) + x_t - D_t(x_t;\theta) \right\|^2$
    Energy Pinning$(\eta) = \nabla_\eta \, \mathbb{E}_x \|(-U_{t=1}(x;\eta)) - \beta_{i+1} \log \pi(x)\|^2$
    $\theta \leftarrow$ `FirstOrderOptimizer`$(\theta, \text{Score Matching}(\theta), \text{Target Score}(\theta))$
    $\eta \leftarrow$ `FirstOrderOptimizer`$(\eta, \text{Energy Matching}(\eta), \text{Energy Pinning}(\eta))$
  end for
  return trained parameters θ^*, η^*

To train the energy model $U_t(x;\eta)$, which plays the central role in the inference-time annealing (see Section 3.1), we follow (Thornton et al., 2025) and distill the learned score model to the parametric energy model via the following regression loss (w.r.t. η), i.e.

$$\text{EBM Distillation}(\eta) = \mathbb{E}_{t,x_t,x} \lambda(t) \left\| \sigma_t^2 (-\nabla_{x_t} U_t(x_t;\eta)) + x_t - D_t(x_t;\theta) \right\|^2, \tag{15}$$

where, the expectation is taken w.r.t. the same random variables as in Eq. (13). Note that, in contrast to the denoising score matching loss in Eq. (13), the "target" in Eq. (15) does not depend on x, which means that its variance for the same x_t is zero, stabilizing the training of the energy based model $U_t(x;\eta)$.

Finally, to use all the supervision signal available in the problem, we use the target unnormalized density $\pi(x)^{\beta_{i+1}}$ as the regression target for the end-point energy-based model $U_{t=1}(x;\eta)$, and introduce the following loss

$$\text{Energy Pinning}(\eta) = \mathbb{E}_{\pi^{\beta_{i+1}}(x)} \|(-U_{t=1}(x;\eta)) - \beta_{i+1} \log \pi(x)\|^2. \tag{16}$$

Notably, this loss allows for fixing the gauge present due to the shift invariance of the energy-based model $(\nabla_x(U_t(x;\eta)) = \nabla_x(U_t(x;\eta) + \text{const}))$. In practice, we observe that this loss significantly stabilizes the training and improves the final performance. We present the pseudo-code for the training loop in Algorithm 1, where we simultaneously optimize all the introduced loss functions to train a diffusion model at temperature $1/\beta_{i+1}$. In practice, we find that sequential training of models demonstrates the best performance. Furthermore, we initialize the model for the next temperature $1/\beta_{i+1}$ with the parameters of the trained model for the temperature $1/\beta_i$.

4. Related work

Diffusion-based Sampling. A variety of amortized samplers that use properties of diffusion models have recently been proposed in the literature. Simulation-based approaches that also exploit the fast mode-mixing of diffusion models include Berner et al. (2022); Vargas et al. (2023); Zhang & Chen (2022); Richter et al. (2024); Vargas et al. (2024), which exploit diffusion processes for fast mode mixing. Conversely, simulation-free methods like iDEM (Akhound-Sadegh et al., 2024), SB with Föllmer drift (Huang et al., 2021), and TSM (De Bortoli et al., 2024) offer more scalable approaches to learning the score but suffer from inefficient and high variance score estimates far from the data. Finally, new diffusion bridges have also risen to prominence with underdamped dynamics (Blessing et al., 2024) and bridges with SMC (Chen et al., 2024).

Inference-time Resampling. The inference-time annealing scheme proposed in Proposition 1 connects several recently proposed methods. Namely, for $\xi_t \equiv 0$, it closely matches the importance sampling of the continuous normalizing flows proposed in (Köhler et al., 2020). Indeed, Eq. (11) becomes a probability flow ODE, and Eq. (12) becomes an integration of the log-density-ratio, where the target density can be defined either as a linear interpolation of log-densities or only in the final point as the target density. Furthermore, Proposition 1 is an application of the Feynman-Kac formula to annealing and non-equilibrium dynamics simultaneously. Indeed, for $\gamma = 1$, this proposition becomes the result proposed in (Vaikuntanathan & Jarzynski, 2008; Albergo & Vanden-Eijnden, 2024); whereas, for $s_t(x) = -\nabla U_t(x;\eta) = \nabla \log p_t(x)$, this proposition becomes the result from (Skreta et al., 2025). In practice, however, these equalities are not satisfied because we use learned models for the vector field $s_t(x) = s_t(x;\theta)$ and the energy-based model $U_t(x;\eta)$.

5. Experiments

We evaluate PITA on molecular conformation sampling tasks including toy Lennard-Jones system of 13 particles (LJ-13) and Alanine peptide systems of varying sizes (Alanine Dipeptide and Tripeptide) in Cartesian coordinate space. Throughout, we assume the access to a short MCMC chain run at high temperature. Note that we do not require these chains to be well mixed, but only require them to cover the modes, a much less stringent requirement (See Appendix D). For metrics, we use sample-based metrics such as 2-Wasserstein distance on Ramachandran coordinates ($\mathbb{T}$-$\mathbb{W}_2$) and energy distribution ($\mathcal{E}$-$\mathbb{W}_1$, $\mathcal{E}$-$\mathbb{W}_2$), to assess mode coverage and precision respectively. We also compare the KL divergence between the Ramachandran plots of the ground-truth MD samples and the generated samples (RAM-KL), as well as Wasserstein distances on the first two TICA (Time-lagged Independent Component Analysis) coordinates of ground-truth and generated samples.

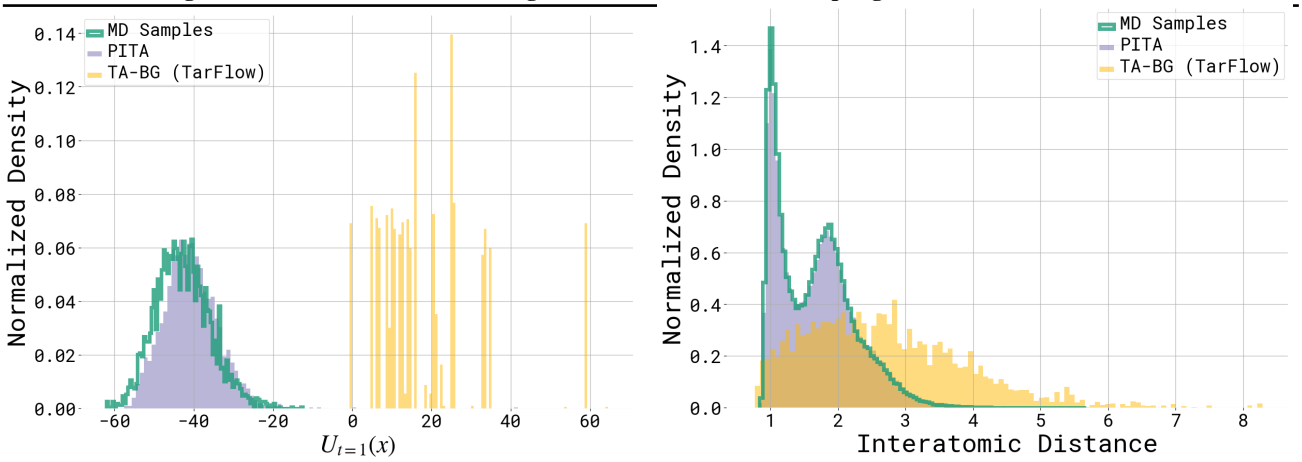
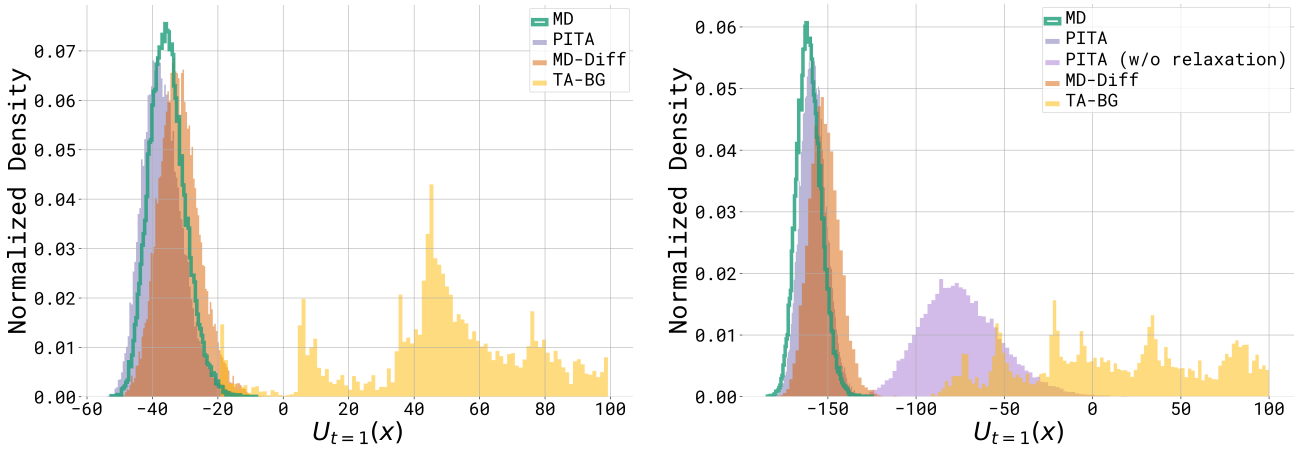


Figure 2: LJ-13 sampling task. We compare the distribution of the interatomic distances and energy of the particles in the MCMC dataset(ground-truth), samples generated using a PITA model, and TA-BG progressively trained from high temperature to sample from the target distribution.



(a) Alanine Dipeptide

(b) Alanine Tripeptide

Figure 3: Molecular conformation sampling tasks. We compare the energy distribution of the ground-truth MD dataset and the samples generated using different models at 300K. We use 30k samples for the plots.

Finally, we report the computational expense of all methods using the total number of energy evaluation calls.

Baselines. We compare PITA with three different baselines: Temperature Annealed Boltzmann Generators (TA-BG, (Schopmans & Friederich, 2025)), diffusion model trained on molecular simulation (MD) data collected at the target temperature (MD-Diff), and importance sampling using continuous normalizing flows (Köhler et al., 2020). For the LJ-13 dataset, we additionally compare the performance of the model with two state-of-the-art diffusion-based sampling algorithms: namely, iDEM (Akhound-Sadegh et al., 2024) and (Havens et al., 2025)); however, as none of the current diffusion-based approaches are able to achieve competitive performance on the small protein tasks, we only compare to the TA-BG baseline for those datasets. Further experimental details of the baselines are provided in the Appendix G.

Architecture. In training PITA, we use EGNN (Satorras et al., 2021) as the model backbone for LJ-13, and DiT(Peebles & Xie, 2023) for ALDP and AL3. Our training follows a sequential temperature schedule, proceeding from high to low temperatures. After training at a given temperature for a fixed number of epochs, we generate samples at the next lower temperature and continue training at that temperature. For LJ-13, we train a single diffusion model conditioning it on β and using the data across all previously seen temperatures. For molecular conformation sampling tasks, we adopt a fine-tuning approach, where at each temperature step, the model is trained only on the newly generated samples corresponding to the current temperature without revisiting earlier ones. For the TA-BG baseline, we train TarFlow (Zhai et al., 2024) with adaptations suited to molecular data for all three systems, and for MD-Diff, we use the same DiT architecture we used in PITA. We parameterize the energy network using the parameterization in Thornton et al. (2025) and use the preconditioning (c_s, c_{out}, c_{in}, c_t) of (Karras et al., 2022) for both energy and score networks. Further training details and hyperparameters are provided in Appendix G.

Hyperparameters. Proposition 1 allows for many choices of the vector field $s_t(x)$. In practice, we set it proportional to the score model $s_t(x) \propto s_t(x;\theta)$ and try several scaling coefficients (see Appendix E). Finally, one can easily add the time-

Table 1: Performance of methods for the ALDP sampling task. The starting temperature is $T_L = 1200$ K, annealed to target $T_S = 300$ K. Metrics are calculated over 10k samples and standard deviations over 3 seeds.

Model	Rama-KL	Tica-$\mathcal{W}_1 \downarrow$	Tica-$\mathcal{W}_2 \downarrow$	Energy-$\mathcal{W}_1 \downarrow$	Energy-$\mathcal{W}_2 \downarrow$	$\mathbb{T}$-$\mathcal{W}_2$	#Energy Evals
PITA	4.773 ± 0.460	$\mathbf{0.112} \pm \mathbf{0.006}$	$\mathbf{0.379} \pm \mathbf{0.028}$	$\mathbf{1.530} \pm \mathbf{0.068}$	$\mathbf{1.615} \pm \mathbf{0.053}$	$\mathbf{0.270} \pm \mathbf{0.023}$	5×10^7
MD-Diff	$\mathbf{1.308} \pm \mathbf{0.072}$	0.113 ± 0.001	0.579 ± 0.004	3.627 ± 0.023	3.704 ± 0.026	0.310 ± 0.001	5×10^7
TA-BG	14.993 ± 0.002	0.219 ± 0.013	0.685 ± 0.034	83.457 ± 0.070	86.176 ± 0.104	0.979 ± 0.012	5×10^7

Table 2: Performance of methods for the AL3 sampling task. The starting temperature is $T_L = 1200$ K, annealed to target $T_S = 300$K. Metrics are calculated over 10k samples and standard deviations over 3 seeds.

Model	Rama-KL	Tica-$\mathcal{W}_1 \downarrow$	Tica-$\mathcal{W}_2 \downarrow$	Energy-$\mathcal{W}_1 \downarrow$	Energy-$\mathcal{W}_2 \downarrow$	$\mathbb{T}$-$\mathcal{W}_2$	#Energy Evals
PITA	$\mathbf{1.209} \pm \mathbf{0.144}$	0.272 ± 0.017	0.952 ± 0.055	$\mathbf{2.567} \pm \mathbf{0.108}$	$\mathbf{2.592} \pm \mathbf{0.107}$	0.521 ± 0.006	8×10^7
PITA (w/o relaxation)	8.535 ± 0.254	0.405 ± 0.014	0.999 ± 0.043	86.270 ± 0.294	87.695 ± 0.294	0.651 ± 0.013	5×10^7
MD-Diff	7.262 ± 2.278	$\mathbf{0.106} \pm \mathbf{0.008}$	$\mathbf{0.525} \pm \mathbf{0.025}$	8.745 ± 0.033	8.949 ± 0.120	$\mathbf{0.487} \pm \mathbf{0.013}$	5×10^7
TA-BG	14.988 ± 0.009	0.321 ± 0.001	0.648 ± 0.000	173.042 ± 0.717	178.558 ± 0.732	1.310 ± 0.004	5×10^7

dependent schedule γ_t by adding the extra term to the weights, i.e. $d \log w_t =$ Eq. (12) $- U_t(x_t; \eta)\partial\gamma_t/\partial t dt$, which we study in Appendix E. We use the noise schedule from (Karras et al., 2022), where, for all experiments, we set $\sigma_{\max} = 80$ and $\rho = 7$. For LJ-13 and molecular conformation sampling tasks, we use $\sigma_{\min} = 0.05, 0.01$, respectively.

5.1. Main results

LJ-13. We first consider a Lennard-Jones (LJ) system of 13 particles to demonstrate the effectiveness of training a sampler at a high temperature ($T_L = 4$), followed by annealing to a lower temperature ($T_S = 1$). As shown in Table 3, we compare the performance of PITA with existing baselines. A visual comparison to TA-BG is provided in Fig. 2. We evaluate each method using the 2-Wasserstein metric over interatomic distance and energy distributions. To ensure consistency, we exclude samples with energy above 1000 across all methods; this notably impacts TA-BG, removing approximately 60% of its samples. Even under this filtering, PITA consistently outperforms TA-BG and other baselines trained directly at the target temperature.

Table 3: LJ-13 sampling task. The starting temperature is $T_L = 4$, annealed to $T_S = 1$.

Algorithm	Distance-$\mathcal{W}_2 \downarrow$	Energy-$\mathcal{W}_2 \downarrow$
iDEM	1.61 ± 0.01	30.78 ± 24.46
Adjoint Sampling	1.67 ± 0.01	2.40 ± 1.25
TA-BG (TarFlow)	1.21 ± 0.02	61.47 ± 0.12
PITA (Ours)	**0.04 ± 0.00**	**2.26 ± 0.21**

Alanine Dipeptide. We apply PITA to the task of sampling Alanine Dipeptide at a target temperature of $T_S = 300$ K, given initial samples at a higher temperature of $T_L = 1200$ K. We use annealing steps of 1200 K, 755.95 K, 555.52 K, 300 K. These temperatures correspond to a subset of the temperatures from (Schopmans & Friederich, 2025), as PITA does not require as many annealing steps to achieve competitive performance. We also analyze the performance of the model, taking larger annealing steps in Appendix E. As shown in Table 1, PITA consistently outperforms both the diffusion-based baseline and TA-BG across all evaluation metrics, achieving a particularly large margin in energy-related metrics. We further present TICA plots of the generated samples at the target temperature in Fig. 4. PITA successfully recovers the essential slow collective dynamical modes of the system, which baseline methods fail to capture. Additionally, we find that while TA-BG performs reasonably well at earlier stages of training at higher temperatures, its performance deteriorates as temperature decreases. Such decline is likely due to the increasing difficulty in generating high-quality proposals as the temperature decreases, which is crucial in the importance sampling used for subsequent training stages. Additional details on training dynamics across all temperatures for PITA and TA-BG can be found in Appendix F.

Alanine Tripeptide. We further evaluate the performance of PITA on a larger molecular system, Alanine Tripeptide (AL3). We employ a temperature annealing schedule with intermediate steps at 1200 K, 755.95 K, 555.52 K, 408.24 K, 300 K. As shown in Fig. 5, PITA again successfully recovers the essential dynamical modes of the system, indicating its capability of generating samples that align with the dominant kinetic features of the underlying dynamics. In practice, we also observe that performing a short additional MD refinement at the target temperature (300 K) after generation further improves the physical plausibility and smoothness of the trajectories, leading to more accurate estimates of the free energy landscape. In Table 2, we provide quantitative analysis of the performance of the models. Notably, despite resulting in a better mode coverage, PITA performs worse than baselines according to Tica$-\mathcal{W}_1$ and Tica$-\mathcal{W}_2$, which suggest that it does not fully recover the correct relative weights of the modes.

6. Conclusion

In this paper, we propose PITA a new framework to train diffusion-based samplers by introducing two mechanisms of interpolating a target Boltzmann density by changing the temperature and defining a diffusion noising process. We demonstrated that PITA allows the progressive training of a sequence of diffusion models that go from high temperature, where ground truth data is simple to collect, to the lower temperature target temperature. Using PITA we demonstrated equilibrium sampling of N-body particle systems, and, for the first time, equilibrium sampling of alanine dipeptide and tripeptide in Cartesian coordinates. Importantly, we demonstrate PITA requires drastically fewer energy evaluations than existing diffusion samplers.

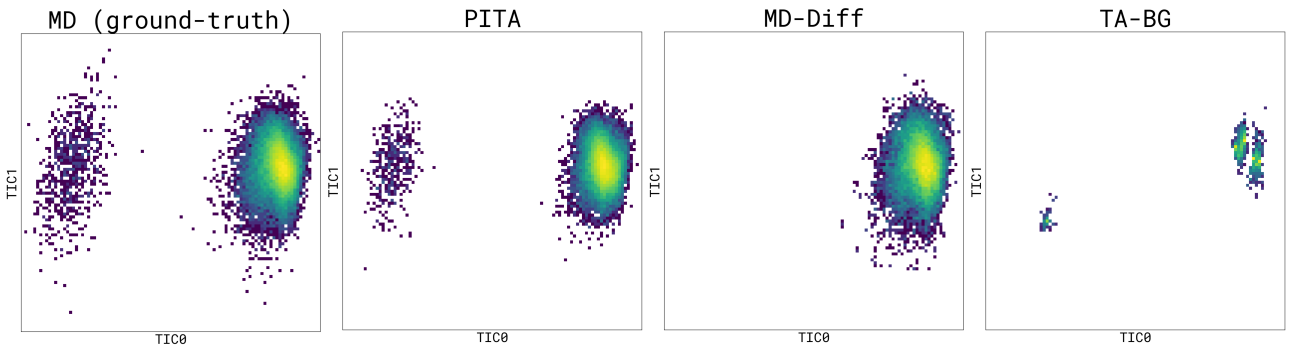


Figure 4: TICA plots for Alanine Dipeptide (ALDP) at 300K obtained from different methods using 30k samples. Each panel shows the free energy landscape along the top two TICA components which capture the dominant slow transitions in the system.

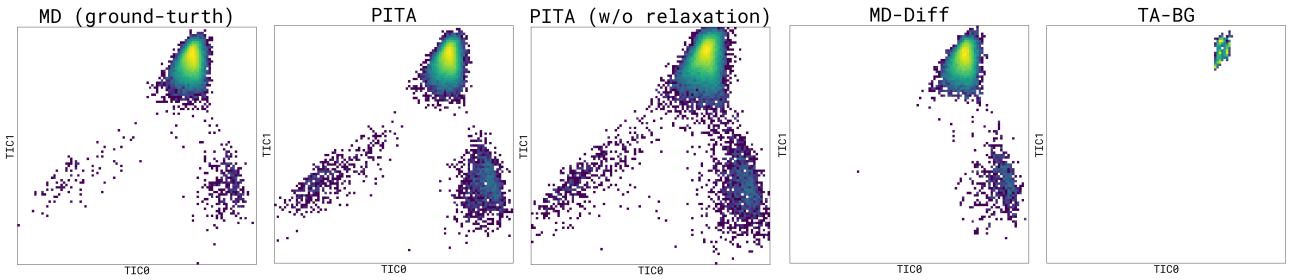


Figure 5: TICA plots for Alanine Tripeptide (AL3) at 300K obtained from different methods using 30k samples.

We believe that PITA represents a step forward in the scalability of diffusion-based samplers and opens up ripe avenues for future work including improved training strategies and regimes for energy-based models that are in service of the PITA framework. Another natural direction for future work includes automatically determining the optimal temperature jump when instantiating our Feynman-Kac PDE to generate asymptotically unbiased samples at lower temperatures.

Limitations. To obtain a consistent estimator or an importance sampling scheme, one has to define a density model of the generated samples. For this, PITA relies on training an additional energy-based model, which is a notoriously challenging task. Furthermore, simultaneous training and inference of both the score model and the energy-based model introduces additional computational and memory requirements.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Akhound-Sadegh, T., Rector-Brooks, J., Bose, A. J., Mittal, S., Lemos, P., Liu, C.-H., Sendera, M., Ravanbakhsh, S., Gidel, G., Bengio, Y., Malkin, N., and Tong, A. Iterated denoising energy matching for sampling from Boltzmann densities. In *International Conference on Machine Learning*, 2024.

Albergo, M. S. and Vanden-Eijnden, E. Nets: A non-equilibrium transport sampler. *arXiv preprint arXiv:2410.02711*, 2024.

Berner, J., Richter, L., and Ullrich, K. An optimal control perspective on diffusion-based generative modeling. *arXiv preprint arXiv:2211.01364*, 2022.

Blessing, D., Jia, X., Esslinger, J., Vargas, F., and Neumann, G. Beyond ELBOs: A large-scale evaluation of variational methods for sampling. In *International Conference on Machine Learning*, 2024.

Chen, J., Richter, L., Berner, J., Blessing, D., Neumann, G., and Anandkumar, A. Sequential controlled Langevin diffusions. *arXiv preprint arXiv:2412.07081*, 2024.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.

De Bortoli, V., Hutchinson, M., Wirnsberger, P., and Doucet, A. Target score matching. *arXiv preprint arXiv:2402.08667*, 2024.

Del Moral, P., Doucet, A., and Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68 (3):411–436, 2006.

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. S. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023.

Frenkel, D. and Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*. Elsevier, 2023.

Havens, A., Miller, B. K., Yan, B., Domingo-Enrich, C., Sriram, A., Wood, B., Levine, D., Hu, B., Amos, B., Karrer, B., et al. Adjoint sampling: Highly scalable diffusion samplers via adjoint matching. *arXiv preprint arXiv:2504.11713*, 2025.

He, J., Du, Y., Vargas, F., Zhang, D., Padhy, S., OuYang, R., Gomes, C., and Hernández-Lobato, J. M. No trick, no treat: Pursuits and challenges towards simulation-free training of neural samplers. *arXiv preprint arXiv:2502.06685*, 2025.

Hénin, J., Lelièvre, T., Shirts, M. R., Valsson, O., and Delemotte, L. Enhanced sampling methods for molecular dynamics simulations. *Living Journal of Computational Molecular Science*, 4 (1), 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Huang, J., Jiao, Y., Kang, L., Liao, X., Liu, J., and Liu, Y. Schrödinger-Föllmer sampler: sampling without ergodicity. *arXiv preprint arXiv:2106.10880*, 2021.

Jarzynski, C. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2963, 1997.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Köhler, J., Klein, L., and Noé, F. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International Conference on Machine Learning*, 2020.

Leimkuhler, B. and Matthews, C. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Springer, 2015.

Liu, J. S. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

Neal, R. M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Neklyudov, K., Brekelmans, R., Severo, D., and Makhzani, A. Action matching: Learning stochastic dynamics from samples. In *International Conference on Machine Learning*, 2023.

Ohno, K., Esfarjani, K., and Kawazoe, Y. *Computational Materials Science: From Ab Initio to Monte Carlo Methods*. Springer, 2018.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Richter, L., Berner, J., and Liu, G.-H. Improved sampling via learned diffusions. *International Conference on Learning Representations*, 2024.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*, 2021.

Schopmans, H. and Friederich, P. Temperature-annealed Boltzmann generators. *arXiv preprint arXiv:2501.19077*, 2025.

Skreta, M., Akhound-Sadegh, T., Ohanesian, V., Bondesan, R., Aspuru-Guzik, A., Doucet, A., Brekelmans, R., Tong, A., and Neklyudov, K. Feynman–Kac correctors in diffusion: Annealing, guidance, and product of experts. In *International Conference on Machine Learning*, 2025.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Stoltz, G., Rousset, M., and Lelièvre, T. *Free Energy Computations: A Mathematical Perspective*. World Scientific, 2010.

Swendsen, R. H. and Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607, 1986.

Thornton, J., Bethune, L., Zhang, R., Bradley, A., Nakkiran, P., and Zhai, S. Composition and control with distilled energy diffusion models and sequential Monte Carlo. *arXiv preprint arXiv:2502.12786*, 2025.

Vaikuntanathan, S. and Jarzynski, C. Escorted free energy simulations: Improving convergence by reducing dissipation. *Physical Review Letters*, 100(19):190601, 2008.

Vargas, F., Grathwohl, W. S., and Doucet, A. Denoising diffusion samplers. In *International Conference on Learning Representations*, 2023.

Vargas, F., Padhy, S., Blessing, D., and Nüsken, N. Transport meets variational inference: Controlled Monte Carlo diffusions. In *International Conference on Learning Representations*, 2024.

Woodard, D., Schmidler, S., and Huber, M. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 2009.

Zhai, S., Zhang, R., Nakkiran, P., Berthelot, D., Gu, J., Zheng, H., Chen, T., Bautista, M. A., Jaitly, N., and Susskind, J. Normalizing flows are capable generative models. *arXiv preprint arXiv:2412.06329*, 2024.

Zhang, Q. and Chen, Y. Path integral sampler: a stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022.

A. Proofs

A.1. Inference-Time Annealing

Proposition 1. [Inference-time Annealing] *Annealed density of the energy-based model $q_t(x) \propto \exp(-\gamma U_t(x;\eta))$ matches the marginal densities of the following SDE*

$$dx_t = \Big(-a_t x_t + \frac{\zeta_t^2}{2}(s_t(x_t) - \gamma\xi_t \nabla U_t(x_t;\eta))\Big)dt + \zeta_t\sqrt{\xi_t}dW_t\,, \quad x_0 \sim q_{t=0}(x) \tag{11}$$

$$d\log w_t = \Big[\frac{\zeta_t^2}{2}\langle\nabla, s_t(x_t)\rangle - \gamma\Big\langle \nabla U_t(x_t;\eta), -a_t x_t + \frac{\zeta_t^2}{2}s_t(x_t)\Big\rangle - \gamma\frac{\partial U_t(x_t;\eta)}{\partial t}\Big]dt\,, \tag{12}$$

where $s_t(x)$ is any vector field, a_t, ζ_t, ξ_t are analogous to parameters from Eq. (4), and the sample weights w_t correspond to the SNIS estimator of the Feynman-Kac formula in Eq. (9).

Proof. For the Energy-Based Model $U_t(x;\eta)$, we denote the corresponding Boltzmann density as

$$q_t(x) \propto \exp(-\gamma U_t(x;\eta))\,, \tag{17}$$

where γ is the target inverse temperature. Taking the time-derivative of $q_t(x)$ we get the following equation

$$\frac{\partial q_t(x)}{\partial t} = q_t(x)\Big(-\gamma\frac{\partial U_t(x;\eta)}{\partial t} + \mathbb{E}_{q_t(x)}\gamma\frac{\partial U_t(x;\eta)}{\partial t}\Big)\,, \tag{18}$$

which can be simulated by reweighting the samples from $q_{t=0}(x) \approx \mathcal{N}(0,1)$ according to the following weights

$$w(x_1) = \frac{\exp(-\gamma\int_0^1 dt\; \partial U_t(x_t;\eta)/\partial t)}{\mathbb{E}_{q_0(x_0)}\exp(-\gamma\int_0^1 dt\; \partial U_t(x_t;\eta)/\partial t)}\,. \tag{19}$$

Although this reasoning is theoretically justified, in practice, the variance of this importance-weighted estimate (or resampled distribution) is prohibitively large. That is why one has to introduce additional terms that move samples around tracing the original diffusion process. Namely, we consider the following PDE

$$\begin{aligned}\frac{\partial q_t(x)}{\partial t} &= \pm\Big\langle\nabla, q_t(x)(-a_t x + \frac{\zeta_t^2}{2}s_t(x))\Big\rangle + q_t(x)\Big(-\gamma\frac{\partial U_t(x;\eta)}{\partial t} + \mathbb{E}_{q_t(x)}\gamma\frac{\partial U_t(x;\eta)}{\partial t}\Big)\\ &= -\Big\langle\nabla, q_t(x)(-a_t x + \frac{\zeta_t^2}{2}s_t(x))\Big\rangle + q_t(x)\big(g_t(x) - \mathbb{E}_{q_t(x)}g_t(x)\big)\,,\end{aligned} \tag{20}$$

$$g_t(x) = \gamma\langle\nabla U_t(x;\eta), a_t x\rangle - \gamma\frac{\zeta_t^2}{2}\langle\nabla U_t(x;\eta), s_t(x)\rangle - \langle\nabla, a_t x\rangle + \tag{21}$$

$$+\frac{\zeta_t^2}{2}\langle\nabla, s_t(x)\rangle - \gamma\frac{\partial U_t(x;\eta)}{\partial t}\,, \tag{22}$$

where the term $\langle\nabla, a_t x\rangle$ does not depend on x and cancels out when in the reweighting term. Furthermore, we can introduce the noise term by adding and subtracting the score $\xi_t(\zeta_t^2/2)\nabla\log q_t(x)$, i.e.

$$\frac{\partial q_t(x)}{\partial t} = -\Big\langle\nabla, q_t(x)(-a_t x + \frac{\zeta_t^2}{2}s_t(x) - \gamma\xi_t\frac{\zeta_t^2}{2}\nabla U_t(x;\eta))\Big\rangle + \xi_t\frac{\zeta_t^2}{2}\Delta q_t(x) + \tag{23}$$

$$+ q_t(x)\big(g_t(x) - \mathbb{E}_{q_t(x)}g_t(x)\big)\,, \tag{24}$$

$$g_t(x) = -\gamma\Big\langle\nabla U_t(x;\eta), -a_t x + \frac{\zeta_t^2}{2}s_t(x)\Big\rangle + \frac{\zeta_t^2}{2}\langle\nabla, s_t(x)\rangle - \gamma\frac{\partial U_t(x;\eta)}{\partial t}\,, \tag{25}$$

which can be simulated as

$$dx_t = \Big(-a_t x_t + \frac{\zeta_t^2}{2}(s_t(x_t) - \gamma\xi_t\nabla U_t(x_t;\eta))\Big)dt + \zeta_t\sqrt{\xi_t}dW_t\,, \tag{26}$$

$$d\log w_t = \Big[-\gamma\Big\langle\nabla U_t(x_t;\eta), -a_t x_t + \frac{\zeta_t^2}{2}s_t(x_t)\Big\rangle + \frac{g_t^2}{2}\langle\nabla, s_t(x_t)\rangle - \gamma\frac{\partial U_t(x_t;\eta)}{\partial t}\Big]dt\,. \tag{27}$$

□

Proposition 2. [Convergence to Diffusion] *For $\gamma = 1$ and perfect models $s_t(x) = -\nabla U_t(x;\eta) = \nabla \log p_t(x)$, the variance of the weights in Proposition 1 becomes zero.*

Proof. Indeed, for $\gamma = 1$, Eq. (25) becomes

$$\begin{aligned} g_t(x) = & -\left\langle \nabla U_t(x;\eta), -a_t x + \frac{\zeta_t^2}{2} s_t(x) \right\rangle + \frac{\zeta_t^2}{2} \langle \nabla, s_t(x) \rangle - \frac{\partial U_t(x;\eta)}{\partial t} \\ = & \left\langle \nabla U_t(x;\eta), a_t x - \frac{\zeta_t^2}{2} s_t(x) \right\rangle - \left\langle \nabla, a_t x_t - \frac{\zeta_t^2}{2} s_t(x) \right\rangle + d a_t - \frac{\partial U_t(x;\eta)}{\partial t} , \end{aligned} \tag{28}$$

where d is the dimensionality of the state-space. For $s_t(x) = -\nabla U_t(x;\eta) = \nabla \log p_t(x)$, we have

$$g_t(x) = -\frac{1}{p_t(x)} \left\langle \nabla, p_t(x) \left(a_t x_t - \frac{\zeta_t^2}{2} \nabla \log p_t(x) \right) \right\rangle + \frac{\partial \log p_t(x)}{\partial t} + \frac{\partial \log Z_t}{\partial t} + d a_t \tag{29}$$

$$= \frac{1}{p_t(x)} \underbrace{\left[-\left\langle \nabla, p_t(x) \left(a_t x - \frac{\zeta_t^2}{2} \nabla \log p_t(x) \right) \right\rangle + \frac{\partial p_t(x)}{\partial t} \right]}_{=0 \text{ due to Eq. (3)}} + \frac{\partial \log Z_t}{\partial t} + d a_t , \tag{30}$$

where $Z_t = \int dx\, p_t(x)$. The term in the brackets equals zero due to Eq. (3) since the ground true marginals $p_t(x)$ are defined as the marginals of the diffusion process. Hence, we have

$$g_t(x) = \frac{\partial \log Z_t}{\partial t} + d a_t = \text{ constant of } x , \tag{31}$$

which becomes zero after the normalization $g_t(x) - \mathbb{E}_{q_t(x)} g_t(x)$, which concludes the proof. □

A.2. Inference-Time Geometric Averaging

For the diffusion process with marginals $p_t(x)$ and the target distribution $p_{t=1}(x) \propto \mathcal{N}(x \,|\, 0, \mathbb{1})^{(1-\beta_i)} \pi(x)^{\beta_i}$, we assume having the energy model $U_t(x;\eta)$ and the score model $s_t(x;\theta)$. Then, for the following density

$$q_t(x) \propto \exp\left(-\frac{\beta_{i+1}}{\beta_i} U_t(x;\eta) - \frac{\beta_{i+1} - \beta_i}{\beta_i} \log \mathcal{N}(x \,|\, 0, (\alpha_{1-t}^2 + \sigma_{1-t}^2)\mathbb{1}) \right) , \tag{32}$$

we have $q_{t=1}(x) \approx \mathcal{N}(x \,|\, 0, \mathbb{1})^{(1-\beta_{i+1})} \pi(x)^{\beta_{i+1}}$. To sample from this density, we derive another SDE that performs inference-time geometric averaging. Analogously to Proposition 1, for $\gamma = 1$ and perfectly trained models, the weights become constant, and this SDE yields the reverse-time diffusion SDE.

Proposition 3. [Inference-time Geometric Averaging] *For the geometric averaging of the energy-based model $q_t(x) \propto \exp\big((1-\gamma)(-U_{t,\beta}(x,\eta)) + \gamma \log \mathcal{N}(x \,|\, 0, \sigma_t^2)\big)$, the weighted samples from $q_t(x)$ can be collected by running the following SDE*

$$\begin{aligned} dx_t = & -a_t x_t + (1-\gamma)\frac{\zeta_t^2}{2}(s_t(x_t) - \xi_t \nabla U_t(x_t;\eta)) - \gamma \frac{1}{\sigma_t^2}\left(1 + \xi_t \frac{\zeta_t^2}{2}\right) x_t + \zeta_t \sqrt{\xi_t} dW_t , \\ d \log w_t = & \left\langle -(1-\gamma)\nabla U_t(x_t;\eta) - \gamma \frac{1}{\sigma_t^2} x_t, -a_t x_t + (1-\gamma)\frac{\zeta_t^2}{2} s_t(x_t) - \gamma \frac{1}{\sigma_t^2} x_t \right\rangle + \qquad (33) \\ & + (1-\gamma)\frac{\zeta_t^2}{2}\langle \nabla, s_t(x_t) \rangle + (1-\gamma)\frac{\partial U_t(x_t;\eta)}{\partial t} + \gamma \frac{1}{\sigma_t^3} \|x_t\|^2 \frac{\partial \sigma_t}{\partial t} . \qquad (34) \end{aligned}$$

where $s_t(x)$ is any vector field. Finally, unweighted samples from $q_t(x)$ can be sampled using SNIS from Eq. (9).

Proof. For the Energy-Based Model $U_t(x;\eta)$, we denote the corresponding geometric averaged density as

$$q_t(x) \propto \exp\big((1-\gamma)(-U_t(x;\eta)) + \gamma \log \mathcal{N}(x \,|\, 0, \sigma_t^2)\big) , \tag{35}$$

where γ is the target inverse temperature. Taking the time-derivative of $q_t(x)$ we get the following equation

$$\frac{\partial q_t(x)}{\partial t} = q_t(x)\big(g_t(x) - \mathbb{E}_{q_t(x)} g_t(x)\big)\,, \tag{36}$$

$$g_t(x) = (1-\gamma)\frac{\partial U_t(x;\eta)}{\partial t} + \gamma \frac{1}{\sigma_t^3}\|x\|^2 \frac{\partial \sigma_t}{\partial t}\,. \tag{37}$$

Assuming that the change of the density is close to the trained diffusion process, we introduce the drift-term corresponding to the score of the marginals

$$\frac{\partial q_t(x)}{\partial t} = \underbrace{\pm\left\langle \nabla, q_t(x)(-a_t x + (1-\gamma)\frac{\zeta_t^2}{2} s_t(x) - \gamma \frac{1}{\sigma_t^2} x)\right\rangle}_{\text{fictitious term}} + q_t(x)\big(g_t(x) - \mathbb{E}_{q_t(x)} g_t(x)\big)\,,$$

$$g_t(x) = (1-\gamma)\frac{\partial U_t(x;\eta)}{\partial t} + \gamma \frac{1}{\sigma_t^3}\|x\|^2 \frac{\partial \sigma_t}{\partial t}\,. \tag{38}$$

Moving the positive term to the weights and interpreting the negative term as the continuity equation, we get

$$\frac{\partial q_t(x)}{\partial t} = -\left\langle \nabla, q_t(x)(-a_t x + (1-\gamma)\frac{\zeta_t^2}{2} s_t(x) - \gamma \frac{1}{\sigma_t^2} x)\right\rangle + q_t(x)\big(g_t(x) - \mathbb{E}_{q_t(x)} g_t(x)\big)\,,$$

$$g_t(x) = \left\langle -(1-\gamma)\nabla U_t(x;\eta) - \gamma\frac{1}{\sigma_t^2}x, -a_t x + (1-\gamma)\frac{\zeta_t^2}{2}s_t(x) - \gamma\frac{1}{\sigma_t^2}x\right\rangle + \tag{39}$$

$$+ (1-\gamma)\frac{\zeta_t^2}{2}\langle \nabla, s_t(x)\rangle + (1-\gamma)\frac{\partial U_t(x;\eta)}{\partial t} + \gamma\frac{1}{\sigma_t^3}\|x\|^2\frac{\partial \sigma_t}{\partial t}\,. \tag{40}$$

Finally, we introduce the noise term by adding the drift

$$\xi_t \frac{\zeta_t^2}{2}\nabla \log q_t(x) = \xi_t \frac{\zeta_t^2}{2}\left(-(1-\gamma)\nabla U_t(x;\eta) - \gamma\frac{1}{\sigma_t^2}x\right)\,. \tag{41}$$

Thus, we get

$$\frac{\partial q_t(x)}{\partial t} = -\left\langle \nabla, q_t(x)\left(-a_t x + (1-\gamma)\frac{\zeta_t^2}{2}(s_t(x) - \xi_t \nabla U_t(x;\eta)) - \gamma\frac{1}{\sigma_t^2}\left(1 + \xi_t\frac{\zeta_t^2}{2}\right)x\right)\right\rangle +$$

$$+ \xi_t\frac{\zeta_t^2}{2}\Delta q_t(x) + q_t(x)\big(g_t(x) - \mathbb{E}_{q_t(x)} g_t(x)\big)\,, \tag{42}$$

$$g_t(x) = \left\langle -(1-\gamma)\nabla U_t(x;\eta) - \gamma\frac{1}{\sigma_t^2}x, -a_t x + (1-\gamma)\frac{\zeta_t^2}{2}s_t(x) - \gamma\frac{1}{\sigma_t^2}x\right\rangle + \tag{43}$$

$$+ (1-\gamma)\frac{\zeta_t^2}{2}\langle \nabla, s_t(x)\rangle + (1-\gamma)\frac{\partial U_t(x;\eta)}{\partial t} + \gamma\frac{1}{\sigma_t^3}\|x\|^2\frac{\partial \sigma_t}{\partial t}\,. \tag{44}$$

The corresponding SDE is

$$dx_t = -a_t x_t + (1-\gamma)\frac{\zeta_t^2}{2}(s_t(x_t) - \xi_t\nabla U_t(x_t;\eta)) - \gamma\frac{1}{\sigma_t^2}\left(1 + \xi_t\frac{\zeta_t^2}{2}\right)x_t + \zeta_t\sqrt{\xi_t}dW_t\,,$$

$$d\log w_t = \left\langle -(1-\gamma)\nabla U_t(x_t;\eta) - \gamma\frac{1}{\sigma_t^2}x_t, -a_t x_t + (1-\gamma)\frac{\zeta_t^2}{2}s_t(x_t) - \gamma\frac{1}{\sigma_t^2}x_t\right\rangle + \tag{45}$$

$$+ (1-\gamma)\frac{\zeta_t^2}{2}\langle \nabla, s_t(x_t)\rangle + (1-\gamma)\frac{\partial U_t(x_t;\eta)}{\partial t} + \gamma\frac{1}{\sigma_t^3}\|x_t\|^2\frac{\partial \sigma_t}{\partial t}\,. \tag{46}$$

□

B. Bridging the Gap at the End-Point

Integrating the dynamics from Propositions 1 and 3 we generate a set of weighted samples $\{(x_{t=1}^k, w_{t=1}^k)\}_{k=1}^K$ that converge to the samples from $q_{t=1}(x)$ when $K \to \infty$. In Section 3.1 we assume that this density is defined as the Boltzmann distribution of the corresponding energy model, i.e. $q_{t=1}(x) \propto \exp(-\beta_{i+1}/\beta_i \cdot U_{t=1}(x;\eta))$, which approximates $\pi^{\beta_{i+1}}$, but does not necessarily match it exactly. Here we describe two possible ways to bridge the gap between the density model and the target density.

The first way to sample from $\pi(x)^{\beta_{i+1}}$ is via Self-Normalized Importance Sampling (SNIS). The integrated weights $w_{t=1}^k = e^{\int_0^1 dt\ g_t(x_t)}$ represent the density ratio between the distribution of the samples $x_{t=1}^k$ and the density of the integrated PDE (see discussion in Section 2.3). Correspondingly, to sample from $\pi(x)^{\beta_{i+1}}$, we have to take into account the density ratio $\pi(x)^{\beta_{i+1}}/q_{t=1}(x)$ and obtain a new estimator, i.e.

$$\mathbb{E}_{\pi(x)^{\beta_{i+1}}}\varphi(x) \propto \mathbb{E}_{q_1(x)}\frac{\pi(x)^{\beta_{i+1}}}{q_1(x)}\varphi(x) \propto \mathbb{E}\left[e^{\int_0^1 dt\ g_t(x_t)}\frac{\pi(x_1)^{\beta_{i+1}}}{q_1(x_1)}\varphi(x_1)\right], \tag{47}$$

$$\mathbb{E}_{\pi(x)^{\beta_{i+1}}}\varphi(x) \approx \sum_{k=1}^{K}\frac{\tilde{w}_1^k}{\sum_{j=1}^n \tilde{w}_1^j}\varphi(x_1^k)\,,\ \tilde{w}_1^k := e^{\int_0^1 dt\ g_t(x_t)}\pi(x_1^k)^{\beta_{i+1}}/q_1(x_1^k)\,, \tag{48}$$

where the new weights $\tilde{w}_1^k$ are obtained from the old ones w_1^k by multiplication with the corresponding density ratio. Note, that the following empirical distribution approximates the target density $\pi(x)^{\beta_{i+1}}$

$$\tilde{\pi}(x)^{\beta_{i+1}} = \sum_{k=1}^{K}\frac{\tilde{w}_1^k}{\sum_{j=1}^n \tilde{w}_1^j}\delta(x - x_1^k)\,. \tag{49}$$

The alternative to importance sampling with the density model proposal is the gradual interpolation between the density model and the target during the integration. In particular, one can satisfy the boundary conditions by defining a smooth interpolant between the boundary densities $p_{t=0} = \mathcal{N}(0, I)$, $p_{t=1} = \pi^{\beta_{i+1}}$ and the annealed density model as follows

$$q_t(x) \propto \exp\Bigg[-\gamma\left(1-\frac{t}{t_1}\right)_+^{\kappa}\log\mathcal{N}(0,1) - \left(1-\frac{1-t}{1-t_2}\right)_+^{\kappa}\beta_{i+1}\log\pi(x) - \tag{50}$$

$$-\left(1-\left(1-\frac{t}{t_1}\right)_+^{\kappa} - \left(1-\frac{1-t}{1-t_2}\right)_+^{\kappa}\right)\frac{\beta_{i+1}}{\beta_i}U_t(x;\eta)\Bigg]\,, \tag{51}$$

where $(x)_+ = \max\{0, x\}$, $0 < t_1 < t_2 < 1$ are the hyperparameters that define switch times between models, and κ is the smoothness parameter. Thus, we guarantee that $q_{t=1}(x) \propto \pi(x)^{\beta_{i+1}}$. However, in practice, we found that this interpolation technique results in a high variance of importance weights.

C. Network Parameterization and Preconditioning

We condition our score network s and our energy network U_t based on findings in EDM (Karras et al., 2022), use an energy parameterization based on Neklyudov et al. (2023) and Thornton et al. (2025), and include a new pre-conditioning on β. All of our networks are based on a backbone $F_\theta(x_t, t, \beta) : (\mathbb{R}^d \times [0, \infty) \times [1, \infty)) \to \mathbb{R}^d$ is a flexible network architecture based on a diffusion transformer (DiT) backbone (Peebles & Xie, 2023). Specifically, we parameterize our denoiser network D_θ as:

$$D_\theta(x_t, t, \beta) := (1 + \beta(c_{skip}(t) - 1)x_t + \beta c_{out}(t)F_\theta(c_{in}(t)x_t, c_{noise}(t)) \tag{52}$$

which allows us to define our score network s_θ as

$$s_\theta(x_t, t, \beta) := \frac{D_\theta(x_t, t, \beta) - x_t}{\sigma_t^2} \tag{53}$$

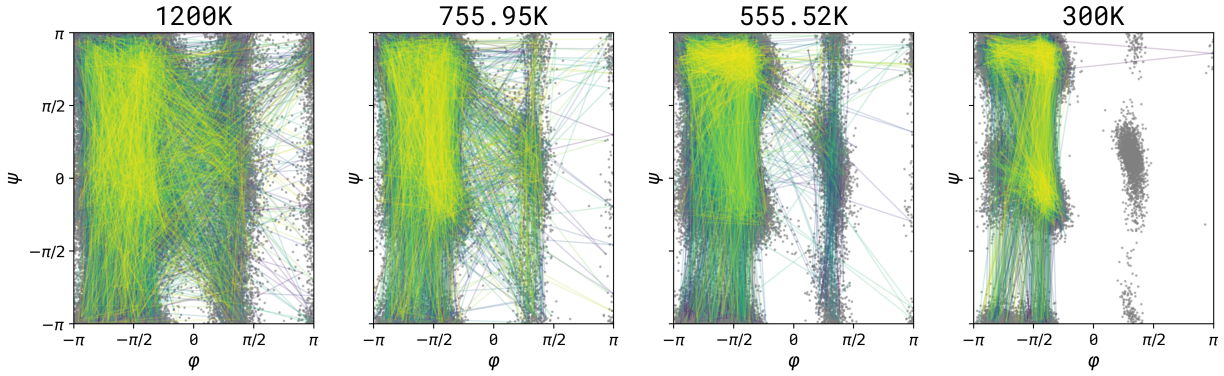
We pre-condition the energy as

$$U_\eta(x_t, t, \beta) := \beta\left(\frac{1 - a_t c_{skip}(t)}{2\sigma_t^2}\|x_t\|^2 - \frac{\xi_t c_{out}(t)}{c_{in}(t)\sigma_t^2}(x_t \cdot F_\eta(c_{in}(t)x_t, c_{noise}(t)))\right) \tag{54}$$

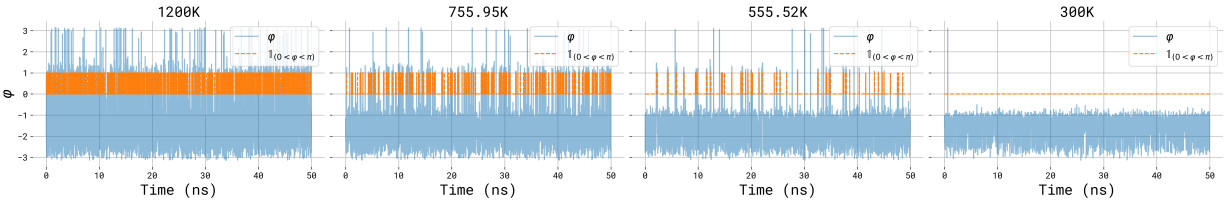
D. Molecular Dynamics Analysis

In Fig. 6 and Fig. 7, we analyze the mixing behaviour of MD simulations for ALDP and AL3 across various annealing temperatures. Specifically, we examine simulations consisting of 50 million steps—matching the quantity of MD data used for training PITA at 1200K. As the temperature decreases, the sampling quality deteriorates: the chains exhibit poorer mixing and fail to explore significant regions of the configuration space, missing major modes of the distribution. This is shown both in Ramachandran and TICA plots, as well as the trace plots of the internal angle ϕ and the second TICA axis. More specifically, for ALDP, we see that the chain switches out of the main mode 5.8%, 3.0%, 1.0% and 0% of the time at temperatures 1200K, 755.95K, 555.52K and 300K, respectively. For AL3, this happens at a rate of 12.7%, 9.2%, 5.8% and 0%.

This motivates training at a higher temperature then annealing to a lower temperature as is done in PITA. As we are able to take advantage of relatively quick mode mixing at higher temperatures and the ability of inference time annealing to recover samples from a lower temperature.

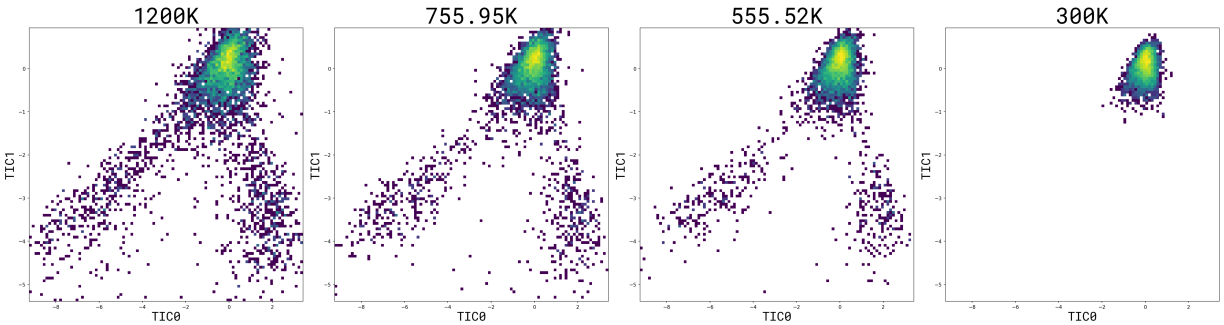


(a) Visualization of the Ramachandran plots of the MD chain over time, the lines are colored from purple to yellow over 50 million MD steps. The gray plot shows the Ramachandran plot of a chain with 1 billion MD steps.

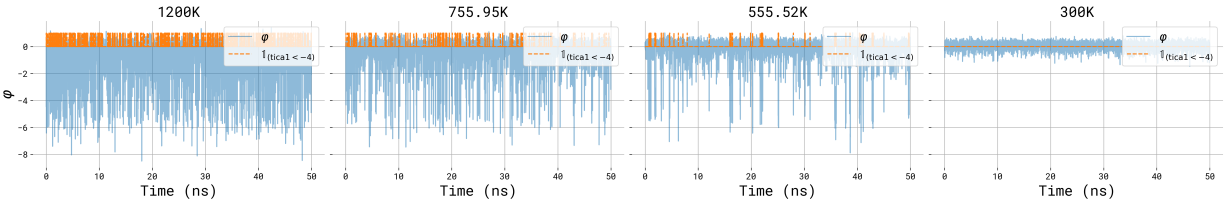


(b) Trace plot for the internal angle ϕ. The blue lines show the value of ϕ across the 50-million-step MD chain. The orange line indicates when the chain switches out of the main modes.

Figure 6: Analysis of the mixing of the MD chains for ALDP, for a 50 million-step MD simulation, across the annealing temperatures.



(a) AL3 TICA plots of ground-truth MD samples across different temperatures, using MD chains of 50 million steps. For all temperatures, the TICA axes are matched to those of 300K.



(b) Trace plot for the second TICA axis of ground-truth MD samples. The blue lines show the value of the second TICA axis across the 50-million-step MD chain. The orange line indicates when the chain switches out of the main mode.

Figure 7: Analysis of the mixing of the MD chains for AL3, for a 50 million-step MD simulation, across the annealing temperatures.

E. Ablation Studies

To evaluate the impact of our design choices, we perform a series of ablation studies examining: (1) the effect of annealing to 300K using different temperature jump sizes, (2) the choice of the γ_t schedule, and (3) the role of resampling and different loss components, as well as comparing out method with a simple classifier-free guidance approach. The temperature jump size ablation is performed for both ALDP and AL3, while the remaining studies are conducted on ALDP.

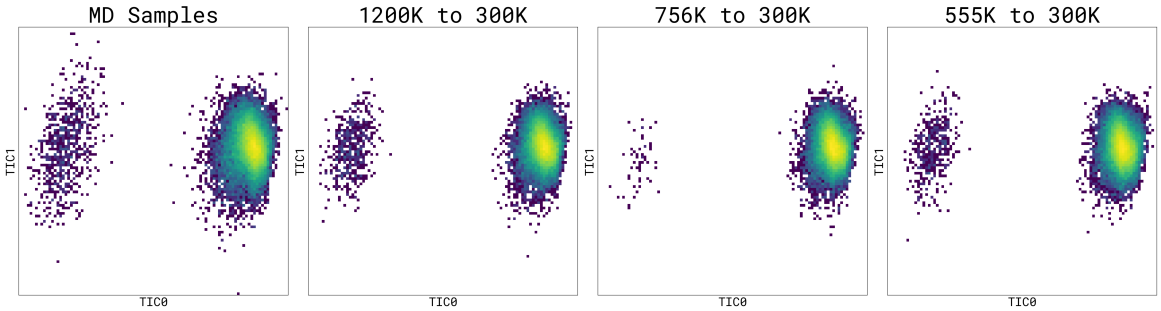
Table 4: Effect of different starting temperatures on annealing performance for ALDP and AL3, evaluated at the final temperature of 300K.

(a) ALDP.

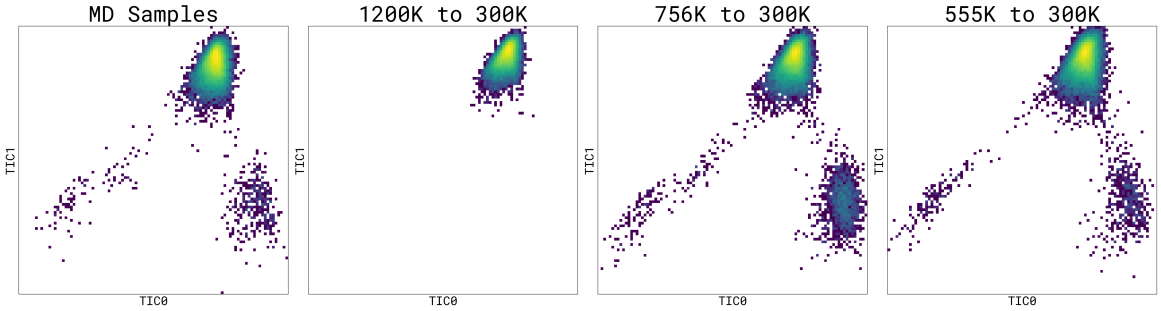
T_L to T_S	Tica-$\mathcal{W}_1 \downarrow$	Tica-$\mathcal{W}_2 \downarrow$	Energy-$\mathcal{W}_1 \downarrow$	Energy-$\mathcal{W}_2 \downarrow$	$\mathbb{T}$-$\mathcal{W}_2 \downarrow$
1200K to 300K	**0.100** $\pm$ **0.004**	**0.297** $\pm$ **0.019**	6.438 $\pm$ 0.024	6.531 $\pm$ 0.021	0.301 $\pm$ 0.023
755.95K to 300K	0.180 $\pm$ 0.002	0.611 $\pm$ 0.003	5.639 $\pm$ 0.072	5.683 $\pm$ 0.070	0.358 $\pm$ 0.018
555.52K to 300K	0.121 $\pm$ 0.004	0.404 $\pm$ 0.019	**1.541** $\pm$ **0.009**	**1.619** $\pm$ **0.010**	**0.270** $\pm$ **0.023**

(b) AL3.

T_L to T_S	Tica-$\mathcal{W}_1 \downarrow$	Tica-$\mathcal{W}_2 \downarrow$	Energy-$\mathcal{W}_1 \downarrow$	Energy-$\mathcal{W}_2 \downarrow$	$\mathbb{T}$-$\mathcal{W}_2 \downarrow$
1200K to 300K	0.291 $\pm$ 0.005	0.558 $\pm$ 0.003	**0.521** $\pm$ **0.122**	**0.597** $\pm$ **0.110**	1.351 $\pm$ 0.014
755.95K to 300K	0.234 $\pm$ 0.009	0.663 $\pm$ 0.019	17.147 $\pm$ 0.105	17.429 $\pm$ 0.107	0.751 $\pm$ 0.006
555.52K to 300K	**0.158** $\pm$ **0.004**	**0.329** $\pm$ **0.025**	40.222 $\pm$ 0.198	40.978 $\pm$ 0.208	**0.621** $\pm$ **0.038**



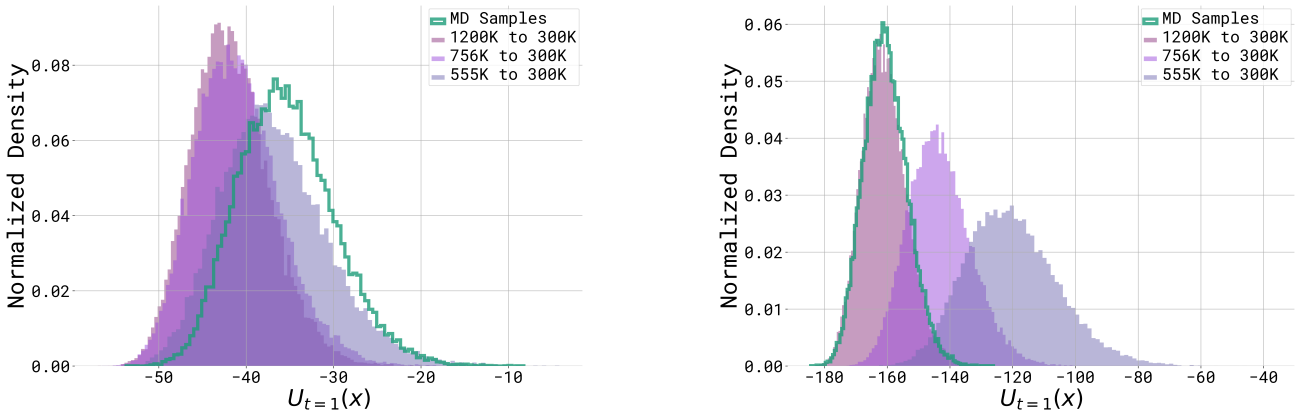
(a) Alanine Dipeptide



(b) Alanine Tripeptide

Figure 8: TICA plot of ALDP and AL3 samples obtained via annealing from various starting temperatures to 300K.

Temperature Jump Sizes in Annealing. To evaluate the effectiveness of the progressive annealing schedule, we compare the performance of models where the system is annealed from different starting temperatures to 300K. For ALDP, skipping intermediate temperatures has the most pronounced impact on energy distribution metrics, as shown in Table 4a and Fig. 9a. In the case of AL3, we find that sequential training is essential for reliably capturing all modes at the lower temperature, as illustrated in Fig. 8b. However, energy metrics degrade, likely due to small deviations in the sampled buffers at each annealing step, which accumulate over time. Accurately capturing the energy distribution in the sampled buffers at each intermediate temperature appears to be more difficult for AL3, which may contribute to the observed degradation. Nonetheless, capturing the correct modes remains a key priority, as modes lost during training are difficult to recover later. In contrast, mild deviations in the energy distribution can often be corrected through short MD relaxation steps, as demonstrated in Section 5.1.



(a) Alanine Dipeptide (b) Alanine Tripeptide

Figure 9: Energy distributions of ALDP and AL3 samples obtained via annealing from various starting temperatures to 300K. The model is trained on ground-truth MD samples at 1200K, and on annealed intermediate samples at 755.95K and 555.52K — the latter being our default training setting.

Table 5: γ_t schedule ablation on ALDP.

γ Schedule	Tica-$\mathcal{W}_1 \downarrow$	Tica-$\mathcal{W}_2 \downarrow$	Energy-$\mathcal{W}_1 \downarrow$	Energy-$\mathcal{W}_2 \downarrow$	$\mathbb{T}$-$\mathcal{W}_2 \downarrow$
Constant	0.115 ± 0.005	0.389 ± 0.014	1.485 ± 0.100	1.584 ± 0.095	$\mathbf{0.258 \pm 0.030}$
Linear	$\mathbf{0.095 \pm 0.009}$	$\mathbf{0.243 \pm 0.048}$	1.453 ± 0.099	1.555 ± 0.099	0.275 ± 0.058
Sigmoid	0.113 ± 0.009	0.339 ± 0.046	$\mathbf{1.443 \pm 0.087}$	$\mathbf{1.550 \pm 0.087}$	0.345 ± 0.027

Table 6: Additional Ablation Results on ALDP.

Model	Tica-$\mathcal{W}_1 \downarrow$	Tica-$\mathcal{W}_2 \downarrow$	Energy-$\mathcal{W}_1 \downarrow$	Energy-$\mathcal{W}_2 \downarrow$	$\mathbb{T}$-$\mathcal{W}_2 \downarrow$
PITA	0.121 ± 0.004	0.404 ± 0.019	$\mathbf{1.541 \pm 0.009}$	$\mathbf{1.619 \pm 0.010}$	0.270 ± 0.023
w/o resampling	0.140 ± 0.007	0.452 ± 0.027	1.606 ± 0.094	1.676 ± 0.075	0.363 ± 0.023
w/o energy pinning loss	$\mathbf{0.098 \pm 0.012}$	$\mathbf{0.291 \pm 0.065}$	4.709 ± 0.091	4.722 ± 0.090	$\mathbf{0.219 \pm 0.021}$
MD-Diff + CFG	0.137 ± 0.007	0.446 ± 0.029	8.106 ± 0.025	8.190 ± 0.026	0.383 ± 0.042

γ_t schedule. We analyze the effects of using different schedules for time-dependent γ_t during inference. In particular, we annneal from 555.52K to 300.0K using a constant schedule, a linear schedule which linearly increases from $\gamma = 1$ to $\gamma = 1.85$, and a sigmoid schedule again, increasing from $\gamma = 1$ to $\gamma = 1.85$. Table 5 shows that the linear schedule generally performs best across the different evaluation metrics. It achieves the lowest values on the TICA metrics, while showing comparable performance to the sigmoid schedule on the energy-based metrics.

Resampling, Energy Pinning Loss, and Classifier Free Guidance. To evaluate the impact of resampling on sample quality, we perform inference from 556K to 300K without applying resampling. Across all metrics, we observe that resampling consistently improves performance. To assess the roles of the energy pinning loss and classifier-free guidance, we retrain a model for the 556K to 300K transition with each component removed. Omitting the energy pinning loss results in a slight improvement in TICA metrics but leads to a noticeable decline in energy metrics, indicating that the loss plays an important role in maintaining accurate energy distributions. Finally, we train a diffusion model on the data generated from PITA at 555.52K, and anneal to 300K simply by scaling the score by γ (similar to classifier-free guidance approaches). This approach shows mixed results, offering no consistent improvement over MD-Diff (which is directly trained on samples at 300K) and performing below the level of PITA in all metrics.

F. Training Dynamics Across Temperatures

In this section, we analyze the performance of the models (PITA and TA-BG) across different temperatures during annealing toward the target temperature on ALDP. Table 7 presents quantitative metrics, demonstrating that PITA consistently achieves lower discrepancies across all temperatures. Additionally, Figure 10 shows the Ramachandran plots at temperatures, further illustrating the ability of the

Table 7: Metrics across temperatures

Temperature	Model	Tica-$\mathcal{W}_1 \downarrow$	Tica-$\mathcal{W}_2 \downarrow$	Energy-$\mathcal{W}_1 \downarrow$	Energy-$\mathcal{W}_2 \downarrow$	$\mathbb{T}$-$\mathcal{W}_2 \downarrow$
755.95K	PITA	**0.024 ± 0.004**	**0.125 ± 0.021**	2.855 ± 0.083	2.886 ± 0.079	**0.134 ± 0.013**
	TA-BG	0.040 ± 0.002	0.178 ± 0.008	**2.065 ± 0.044**	**2.140 ± 0.042**	0.355 ± 0.007
555.52K	PITA	**0.141 ± 0.001**	**0.836 ± 0.004**	**1.420 ± 0.030**	**1.430 ± 0.027**	**0.219 ± 0.015**
	TA-BG	0.337 ± 0.009	0.967 ± 0.013	48.486 ± 0.042	56.897 ± 0.059	1.135 ± 0.004
300K	PITA	**0.112 ± 0.006**	**0.379 ± 0.028**	**1.530 ± 0.068**	**1.615 ± 0.053**	**0.270 ± 0.023**
	TA-BG	0.219 ± 0.013	0.685 ± 0.034	83.457 ± 0.070	86.176 ± 0.104	0.979 ± 0.012

model to generate physically realistic samples that capture the temperature-dependent conformational landscape at each step of the annealing process. TA-BG demonstrates reasonable performance at 755.95K when initialized with ground-truth samples, reflecting its ability to model high-temperature distributions under ideal conditions. However, its performance deteriorates when transitioning to lower temperatures using recursively generated samples for importance sampling, indicated by the mode collapse in the Ramachandran plots, where the conformational diversity sharply diminishes.

G. Additional Experimental Details

G.1. Network Parameters

PITA. For LJ-13, we use equal loss weights for energy pinning, denoising score matching, and EBM distillation. We use the noise schedule of Karras et al. (2022), with the following parameters: $\sigma_{\min} = 0.05$, $\sigma_{\max} = 80$ and $\rho = 7$. The model uses EGNN (Satorras et al., 2021) with approximately 90k parameters, consisting of three layers and a hidden dimension of 32. For ALDP and AL3, the energy pinning, denoising score matching, and EBM distillation components of the loss are weighted equally at 1.0, with an additional target score matching loss weighted at 0.01. We use the same noise schedule as the LJ-13 experiment, using a smaller $\sigma_{\min}$ of 0.01. We use DiT (Peebles & Xie, 2023) comprising six layers and six attention heads, with a hidden size of 192 and a total of roughly 12 million parameters. All models are trained with a learning rate of 1×10^{-3} without any weight decay. For ALDP and AL3, we use Exponential Moving Average (EMA) with a decay rate of 0.999, updating every gradient step.

MD-Diff. We train the diffusion model on the MD trajectories generated at the target temperature. This serves as a strong baseline, since we have direct access to the ground-truth samples, unlike PITA and TA-BG. To ensure a controlled comparison, the length of the MD chain used to train the diffusion model is chosen such that the total number of energy evaluations matches the computational budget used to train PITA over all annealing steps. We provide further analysis on the mixing properties of different lengths of MD chains at low and high temperatures in Appendix D. We use a $\sigma_{\min}$ value of 0.005, while keeping the rest of the model hyperparameters the same as PITA.

TA-BG. In Schopmans & Friederich (2025), TA-BG trains a normalizing flow by minimizing the reverse Kullback–Leibler (KL) divergence at high temperature and progressively refining the model via importance sampling as the temperature is annealed toward the target distribution. We carefully adapt their training pipeline to ensure a consistent and fair comparison. Specifically, we initialize the training with ground-truth MD data rather than learned high-temperature samples, represent molecular configurations in Cartesian rather than internal coordinates, and use the same temperature annealing schedule as PITA. We use TarFlow (Zhai et al., 2024), configured with four meta blocks, each containing four attention layers and a hidden size of 256, resulting in approximately 12 million parameters. We use a learning rate of 1×10^{-4} and employ 60,000 samples at the end of training for each temperature to compute the importance weights used in generating the buffer for the next temperature.

G.2. Metrics

We evaluate model performance using both sample-based metrics and metrics that assess energy distributions. To compare energy distributions between generated samples and ground-truth molecular dynamics (MD) samples, we compute the 1D 1-Wasserstein and 2-Wasserstein distances on the energy histograms. For sample-based evaluation, we measure the 2D wrapped 2-Wasserstein distance of the internal dihedral angles, ϕ and ψ (denoted as $\mathbb{T}$-$\mathcal{W}_2$). Additionally, we calculate the 2D 1-Wasserstein and 2-Wasserstein distances between the first two TICA axes of the ground-truth and generated samples.

G.3. MD Parameters

LJ-13 Parameters. The Lennard-Jones (LJ) potential is an intermolecular potential that models interactions of non-bonding particles. The energy is a function of the interatomic distance of the particles:

$$\mathcal{E}^{\mathrm{LJ}}(x) = \frac{\varepsilon}{2\tau} \sum_{ij} \left(\left(\frac{r_m}{d_{ij}} \right)^6 - \left(\frac{r_m}{d_{ij}} \right)^{12} \right) \tag{55}$$

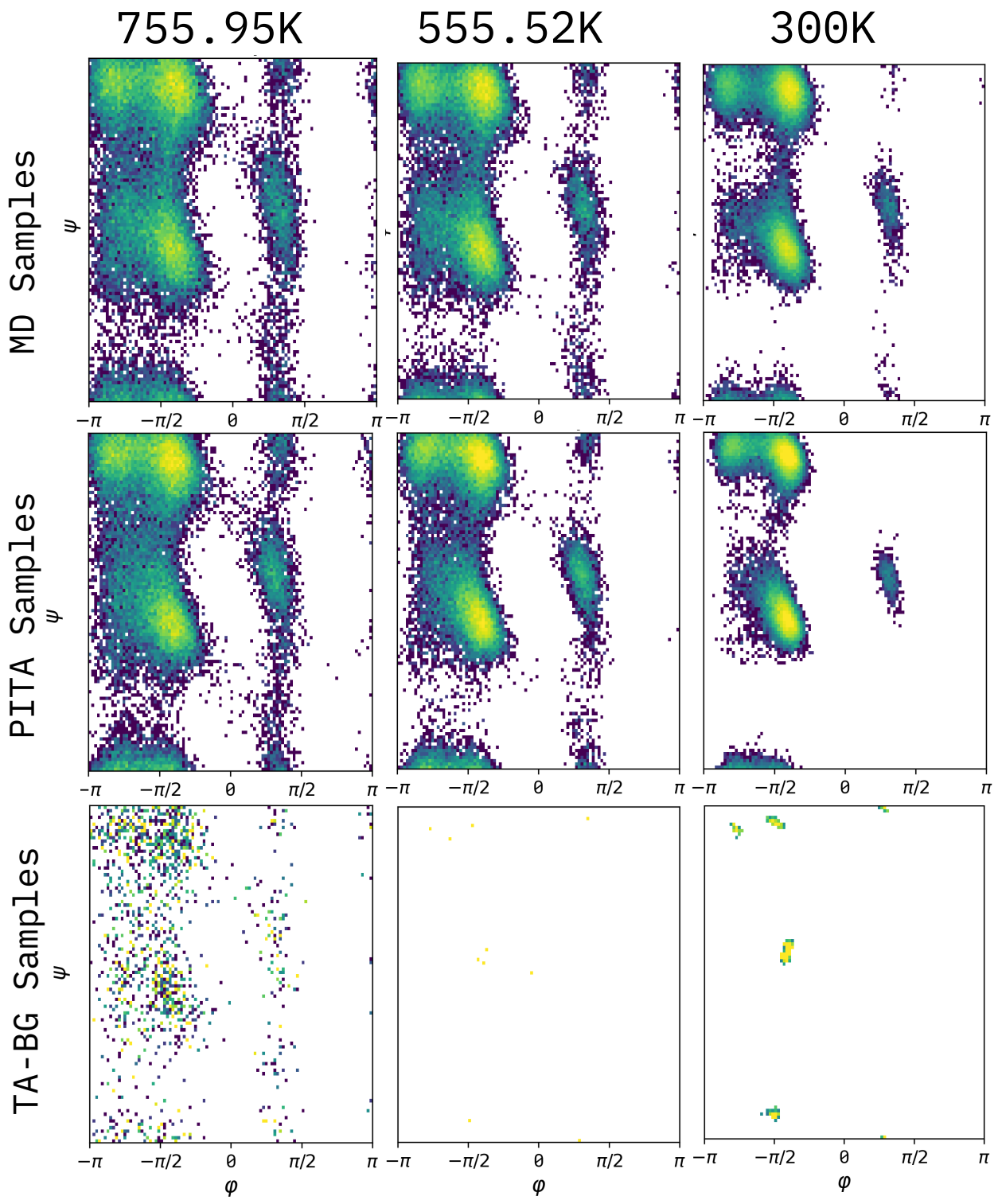


Figure 10: Ramachandran plots for Alanine Dipeptide (ALDP) obtained from different temperatures using 30k samples. We compare the samples from PITA and TA-BG with the ground-truth MD samples.

where the distance between two particles i and j is $d_{ij} = \|x_i - x_j\|_2$ and r_m, τ, ϵ and c are physical constants. As in Köhler et al. (2020), we also add a harmonic potential to the energy so that $\mathcal{E}^{LJ-system} = \mathcal{E}^{\mathrm{LJ}}(x) + c\mathcal{E}^{\mathrm{osc}}(x)$ This harmonic potential is given by:

$$\mathcal{E}^{\mathrm{osc}}(x) = \frac{1}{2}\sum_i ||x_i - x_{\mathrm{COM}}||^2 \tag{56}$$

where x_{COM} is the center of mass of the system. We use $r_m = 1$, $\tau = 1$, $\varepsilon = 2.0$ and $c = 1.0$. For the LJ-13 dataset, we draw MCMC chains using the No-U-Turn-Sampler (NUTS) (Hoffman & Gelman, 2011)

Alanine Parameters. For MD data on ALDP and AL,3, we run two chains one for training and one for test. We use the same simulation parameters for both. For training data we sample shorter chains more frequently (every 100 md steps). To conserve disk space for long test chains, we save every 10k steps. Further parameters can be found in Table 8 and Table 9.

Table 8: `OpenMM` simulation parameters.

Force field	`amber-14`
Integration time step	1 fs
Friction coefficient	$0.3\,\mathrm{ps}^{-1}$
Temperature	310 K
Nonbonded method	`CutoffNonPeriodic`
Nonbonded cutoff	2 nm
Integrator	`LangevinMiddleIntegrator`

Table 9: Training and evaluation dataset parameters.

	Train	Test
Burn-in period	50 ps	50 ps
Sampling interval	0.1 ps	10 ps
Simulation time	50 ns	1 µs

H. Extended Related Work

Annealed Importance Sampling. In the context of AIS (Jarzynski, 1997; Neal, 2001) and parallel tempering (Swendsen & Wang, 1986), our method reduces the number of energy evaluations by learning the models of intermediate marginals. Indeed, when the buffer of samples from the current temperature is sampled, training of the diffusion model does not require new energy evaluations (note that the gradients for target score matching can be cached). Thus, the only time we need to evaluate the energies is for the importance sampling at the final step of the inference-time annealing and for the collection of samples via MCMC at a high temperature. Obviously, for sampling from the target density $\pi(x)$, the trained diffusion model, unlike AIS, allows producing uncorrelated samples without restarting the chain from the prior distribution.

Boltzmann Generators. Noé et al. (2019) proposed training a probabilistic model and resampling the generated samples according to the target Boltzmann density via importance sampling. The entire framework was coined as the Boltzmann Generator. The nature of the probabilistic model can vary as long as it allows for evaluation of the density and efficient training, e.g. continuous normalizing flows (Chen et al., 2018) and the flow matching objective (Lipman et al., 2022) also allow for efficient training and resampling under the Boltzmann generators framework (Köhler et al., 2020; Klein et al., 2023). Boltzmann Generators can also be combined with Annealed Importance Sampling, which allows for the use of SMC style resampling approaches ane enhances their scalability (Tan et al., 2025). However, as we demonstrate empirically, the straightforward resampling with a target density of a different temperature results in high variance of importance weights. Thus, one has to deviate from the Boltzmann Generators framework to perform the inference-time annealing.

Supplement References

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Hoffman, M. D. and Gelman, A. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo, 2011. URL https://arxiv.org/abs/1111.4246.

Jarzynski, C. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2963, 1997.

Klein, L., Krämer, A., and Noé, F. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36:59886–59910, 2023.

Köhler, J., Klein, L., and Noé, F. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International Conference on Machine Learning*, 2020.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Neal, R. M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

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.

Swendsen, R. H. and Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607, 1986.

Tan, C. B., Bose, A. J., Lin, C., Klein, L., Bronstein, M. M., and Tong, A. Scalable equilibrium sampling with sequential boltzmann generators. *arXiv preprint arXiv:2502.18462*, 2025.