Temperature-Regulated Stochastic Sampling for Diffusion-Based High-Quality Molecular Generation

Yuyan Ni[1,2†] **Haoming Kong** [1] **Bo Qiang**[3] **Yanjing Li**[3]
Xin Hong[1] **Shikun Feng**[4 ‡] **Wen Yan**[2] **Yanyan Lan**[1,5 *]
[1] Institute for AI Industry Research (AIR), Tsinghua University
[2] ByteDance Seed
[3] University of Washington
[4]Zhongguancun Institute of Artificial Intelligence, China
[5] Beijing Frontier Research Center for Biological Structure, Tsinghua University

Abstract

Despite the strong performance of diffusion models in molecular generation, the sampling process itself remains largely understudied, with most existing methods adopting off-the-shelf sampling strategies developed for natural images. Yet unlike natural images, molecular distributions are governed by physical laws and are sharply concentrated, making them particularly challenging for standard samplers, which frequently produce invalid or unphysical structures. In this work, we revisit diffusion sampling from a unified stochastic differential equation (SDE) perspective and introduce a general framework parameterized by two interpretable controls: *stochasticity* and *temperature*. Our analysis shows that stochasticity strengthens instantaneous error correction under a bounded score-error condition, while temperature controls distribution sharpness in the low-noise regime. Building on these insights, we develop TREASSURE (Temperature-Regulated Stochastic Sampling), a lightweight, training-free strategy that combines stochasticity scheduling with temperature annealing to better capture the sharply concentrated distributions of molecular data. Extensive experiments on small molecule generation, protein structure prediction, and protein design demonstrate that TREASSURE consistently improves generation quality and produces more physically valid structures. Empirical analyses further corroborate our theoretical findings. The method uses pretrained models without retraining, supporting its practical value across molecular domains.

1 Introduction

Diffusion models [Sohl-Dickstein et al., 2015] have rapidly become a leading generative paradigm for molecular sciences, achieving state-of-the-art results in tasks ranging from small molecule generation [Hoogeboom et al., 2022, Feng et al., 2025, Song et al., 2024a] to protein structure prediction [Abramson et al., 2024, Corley et al., 2025] and de novo protein design [Ingraham et al., 2023, Butcher et al., 2025]. Despite this progress, research has primarily focused on improving model architectures and training objectives, while the sampling process itself, the algorithm that ultimately produces samples from a trained model, has remained comparatively underexplored in the molecular domain.

In diffusion-based generative models, the final generated distribution is significantly affected by the sampling algorithm. In computer vision, extensive studies have demonstrated that sampling strategies

*Corresponding author: Yanyan Lan (lanyanyan@air.tsinghua.edu.cn). † Work done while interning at ByteDance. ‡ Work done while pursuing a PhD at AIR, Tsinghua University.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

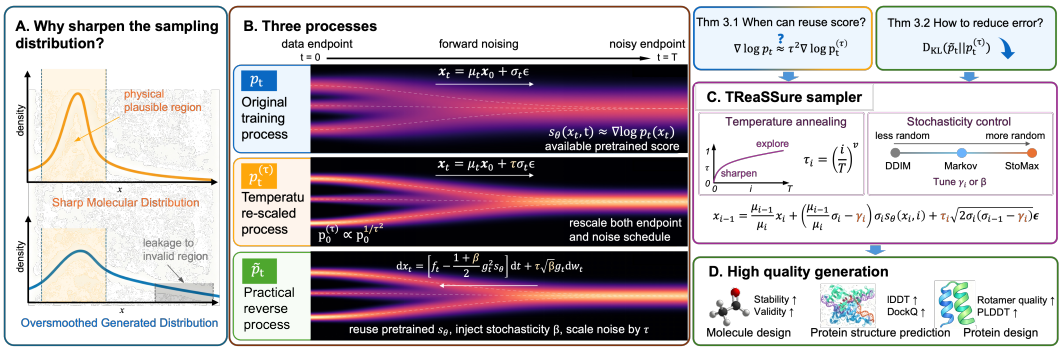


Figure 1: **Overview of TREASSURE. (A)** Molecular distributions are sharply peaked in physically plausible regions, whereas standard samplers over-smooth and leak mass into invalid regions. To address this, **(B)** from a pretrained original process p_t, we introduce a temperature-sharpened reference $p_t^{(\tau)}$ and a practical reverse process $\tilde{p}_t$ that reuses the pretrained score with stochasticity β and noise scaling τ. Theorem 2.1 justifies score reuse for the tempered target and motivates τ annealing; Theorem 2.2 characterizes its instantaneous KL correction under score error. Guided by the analysis, **(C)** the resulting sampler combines temperature annealing with stochasticity control. **(D)** It improves sample quality across molecular generation tasks without retraining.

are far from unique and are important for obtaining good results [Song and Ermon, 2019, Ho et al., 2020, Song et al., 2021a, Zhao et al., 2023]. However, in much of the prior work on molecular diffusion, sampling has been treated as a fixed off-the-shelf component, typically inherited from standard formulations in computer vision. For small-molecule generation, the Equivariant Diffusion Model (EDM) [Hoogeboom et al., 2022] and MiDi [Vignac et al., 2023] follow DDPM [Ho et al., 2020], while GeoBFN [Song et al., 2024a] follows Bayesian flow networks (BFNs) [Graves et al., 2023]. For protein generation, AlphaFold3 [Abramson et al., 2024], RoseTTAFold3 [Corley et al., 2025], and RFdiffusion3 [Butcher et al., 2025] largely adopt samplers inspired by the diffusion design space of Karras et al. [2022], referred to as Karras-EDM below to distinguish it from EDM. A systematic investigation into sampling choices in the molecular domain remains largely absent.

We attribute this performance gap to intrinsic differences between molecular data and natural images. Governed by physical constraints, molecular distributions concentrate sharply around low-energy configurations, while standard samplers produce overly smooth outputs that fail to capture the peaked nature of molecular geometries and often yield invalid high-energy structures, as summarized in Figure 1A. Appendix A and Figure 4 provide direct evidence from bond-angle and bond-length distributions. This failure mode is consistent with neural-network spectral bias [Rahaman et al., 2019]: near the data manifold, the ground-truth score may vary sharply along high-frequency molecular features, whereas the learned score tends to smooth these variations. Because the target distribution is highly concentrated, even modest score errors accumulated along the reverse trajectory can move samples outside physically plausible, high-density regions. This suggests two complementary requirements: sharpening the effective target distribution and correcting accumulated sampling error.

To address these sharpness challenges, we revisit the diffusion sampling process from a stochastic differential equation (SDE) perspective and formalize a unified sampling framework parameterized by two interpretable factors (notation formally defined in Section 2):

$$\mathrm{d}\boldsymbol{x}_t = \left[\frac{\dot{\mu}_t}{\mu_t}\boldsymbol{x}_t - \frac{1+\beta(t)}{2} g_t^2 \,\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right] \mathrm{d}t + \tau\sqrt{\beta(t)}\, g_t \,\mathrm{d}\mathbf{w}_t. \tag{1}$$

(a.) **Stochasticity** ($\beta \geq 0$) governs the relative contribution of the score-driven deterministic drift and the noise-driven exploration along the reverse trajectory. (b.) **Temperature** ($\tau > 0$) reshapes the effective target distribution, regulating its sharpness. Thus τ chooses *which target* is sought, whereas β controls *how it is approached*. These distinct roles are supported by two complementary results. Theorem 2.1 shows that the pretrained score approximates the τ^2-scaled score of the τ-scaled target distribution with $O(\sigma^2)$ error in the low-noise regime, which justifies using a pretrained score to sample from temperature-scaled targets. Moreover, the approximation error decays near the data

manifold, motivating a temperature-annealing schedule. Theorem 2.2 characterizes sampling under an inexact score: below a score-error threshold and for fixed current distributions, larger β accelerates the corrective decay of the KL divergence toward the target. Thus the first result identifies the distribution being pursued, while the second explains when and how stochasticity can more strongly correct accumulated sampling error. Beyond its theoretical guarantees, our framework unifies several existing samplers as special cases, expanding the design space for strategies tailored to the sharp, manifold-localized nature of molecular data.

Based on this analysis, we develop TREASSURE (**T**emperat**ure**-**Re**gul**a**ted **S**tochastic **S**ampling), a lightweight sampler that combines stochasticity scheduling with late-stage temperature annealing. Its discrete formulation spans deterministic, Markov, and maximally stochastic variants, together with their interpolations. TREASSURE operates directly on pretrained diffusion models, requiring neither retraining nor additional model evaluations.

We validate TREASSURE on three molecular generation tasks, observing improvements in the primary physical-validity and structural-quality metrics over standard samplers. For small-molecule generation, our method improves molecular stability by up to 10.0 percentage points and validity by 4.6 percentage points. For protein structure prediction, it yields a $\sim$2% gain in lDDT. For de novo protein design, it enhances side-chain quality and pushes the Pareto-optimal frontier. We further perform ablation studies to examine the empirical implications of our theoretical analysis.

In summary, our key contributions are threefold:
(1) A unified sampling framework is proposed, parameterized by adjustable stochasticity and temperature, explicitly tailored to the sharp, physically constrained nature of molecular distributions. This framework relates existing sampling methods through representative continuous-time and small-step regimes, unlocking a broad design space for molecular sampling.
(2) Rigorous theoretical analysis is derived, providing a quantitative characterization of how stochasticity and temperature affect the sampling distribution and offering practical guidance for use.
(3) Training-free sampling strategies are developed as TREASSURE, which improves the primary quality metrics across small-molecule generation, protein structure prediction, and protein design.

2 A Unified Sampling Framework from an SDE Perspective

We formalize the unified sampling framework previewed in the introduction, with the explicit goal of generating samples from low-temperature distributions. The framework involves three processes and their marginal distributions:
(P1) the original forward noising process with marginal p_t, for which a pretrained score $s_\theta \approx \nabla \log p_t$ is available (defined below);
(P2) an auxiliary temperature-scaled forward process with marginal $p_t^{(\tau)}$, whose ideal reverse score $\nabla \log p_t^{(\tau)}$ is unavailable (constructed in Section 2.1);
(P3) the practical reverse sampling process with marginal $\tilde{p}_t$, which reuses s_θ under stochasticity β and temperature τ (defined below and analyzed in Section 2.3).
Two theorems connect these processes: Theorem 2.1 bounds the error of approximating the τ^2-scaled tempered score of (P2) by the pretrained score from (P1), and Theorem 2.2 characterizes the instantaneous effect of β on the resulting sampling error in the marginal of (P3).

Denote the forward noising process used in diffusion model training (P1) as:

$$\boldsymbol{x}_t = \mu_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N), \tag{2}$$

where $\boldsymbol{x}_0$ and $\boldsymbol{x}_t$ are the clean and corrupted data respectively, and the noise schedule $\mu(t), \sigma(t)$ satisfy conditions including continuously differentiable, non-negative, $\mu(0) = 1$, $\sigma(0) = 0$, $\mu(t)$ non-increasing and $\frac{\mathrm{d}(\sigma(t)/\mu(t))}{\mathrm{d}t} \geq 0$. Let p_t denote the marginal distribution of $\boldsymbol{x}_t$. A neural network $s_\theta(\boldsymbol{x}_t, t)$ is trained to approximate the score function $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$.

Then the proposed family of reverse SDEs (P3) parameterized by two interpretable controls, stochasticity $\beta(t) \geq 0$ and temperature $\tau > 0$, is given by equation 1, where $g_t = \sqrt{2(\sigma_t \dot{\sigma}_t - \sigma_t^2 \frac{\dot{\mu}_t}{\mu_t})}$ and $\mathrm{d}\mathbf{w}_t$ denotes the standard Wiener process.

This formulation recovers widely used sampling schemes as special cases: Setting $\beta = 1, \tau = 1$ recovers the stochastic diffusion SDE; $\beta = 0, \tau = 1$ yields the deterministic diffusion ODE [Song

et al., 2021b]. For $\tau = 1$, the reverse SDE shares the same marginal distribution as the forward process defined by equation 2 for any bounded $\beta(t)$, ensuring exact marginal preservation in principle [Xue et al., 2024a, Karras et al., 2022]. When $\beta \to \infty$, time rescaling $t' = \beta t$ puts the reverse SDE in an annealed Langevin dynamics form [Song and Ermon, 2019], with an additional drift correction scaled by $1/\beta$. For standard Langevin dynamics, scaling the diffusion term by τ produces a stationary distribution proportional to $p(\boldsymbol{x})^{1/\tau^2}$. We provide theoretical analysis to formalize these roles.

2.1 Construction of a Temperature-Scaled Forward Process.

Let the data distribution be $p_0(\boldsymbol{x}) \propto e^{-U(\boldsymbol{x})}$, and define the temperature-scaled target as

$$p_0^{(\tau)}(\boldsymbol{x}) := \frac{1}{Z_\tau} p_0(\boldsymbol{x})^{1/\tau^2} \propto \exp\Big(-\frac{1}{\tau^2} U(\boldsymbol{x})\Big), \tau > 0, \tag{3}$$

where Z_τ is a normalization constant, assumed finite for the temperature under consideration. To analyze temperature control, we introduce the auxiliary process (P2), which tempers both the data distribution and the forward-noise scale:

$$\boldsymbol{x}_t = \mu_t \boldsymbol{x}_0 + \tau \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N), \tag{4}$$

with its SDE representation:

$$\mathrm{d}\boldsymbol{x}_t = \frac{\dot{\mu}_t}{\mu_t} \boldsymbol{x}_t \mathrm{d}t + \tau g_t \mathrm{d}\mathbf{w}_t, \quad g_t = \sqrt{2\Big(\sigma_t \dot{\sigma}_t - \sigma_t^2 \frac{\dot{\mu}_t}{\mu_t}\Big)}. \tag{5}$$

Denote its marginal distribution as $p_t^{(\tau)}$. Unlike training, this auxiliary process tempers both p_0 and the noise scale. Its reverse uses the unavailable $\nabla \log p_t^{(\tau)}$; Theorem 2.1 justifies reusing s_θ by bounding this approximation error; the resulting sampling error is then tracked by Theorem 2.2.

2.2 Theoretical Analysis of Temperature

Prior temperature controls are either heuristic or theoretically justified only under strong assumptions such as Gaussianity as elaborated in Appendix E. Our analysis instead imposes nonparametric smoothness conditions, i.e. globally bounded second- and third-order derivatives, while allowing anisotropic, multimodal, and non-log-concave densities.

Fix $\tau > 0$. Assume that $p_0 \propto e^{-U}$ and $p_0^{(\tau)} \propto e^{-U/\tau^2}$ are normalizable, $U \in C^3(\mathbb{R}^d)$, and $\|\nabla^2 U(x)\|_{\mathrm{op}} \le C_2$, $\|\nabla^3 U(x)\|_{\mathrm{op}} \le C_3$ for all x. Here $\|\cdot\|_{\mathrm{op}}$ denotes the Euclidean induced operator norm for matrices and the third derivative. Assume $\sigma(t) > 0$ for positive times in the interval under consideration and $\mu(t_m) > 0$. For a fixed noisy state $\boldsymbol{x}_t$, define the conditional energy

$$f_{\boldsymbol{x}_t}(\boldsymbol{x}_0) := U(\boldsymbol{x}_0) + \frac{\|\boldsymbol{x}_t - \mu(t)\boldsymbol{x}_0\|^2}{2\sigma(t)^2}. \tag{6}$$

Its Hessian with respect to the clean variable $\boldsymbol{x}_0$ is $\nabla^2_{\boldsymbol{x}_0} f_{\boldsymbol{x}_t}(\boldsymbol{x}_0) = \nabla^2 U(\boldsymbol{x}_0) + \frac{\mu(t)^2}{\sigma(t)^2} I$. Hence, in the low-noise regime, the quadratic noising term dominates any negative curvature of U. This makes $f_{\boldsymbol{x}_t}$ strongly convex even though U itself need not be convex.

Theorem 2.1 (Temperature Scaling Approximation)**.** *Let the noise schedule $\mu(t)$ and $\sigma(t)$ be continuous on $[0,1]$ and positive on $(0,1)$, $\mu(0) = 1$, $\sigma(0) = 0$, with $\mu(t)$ non-increasing and $\sigma(t)/\mu(t)$ non-decreasing on $[0,1]$. For $C_2 > 0$, define $t_m = \sup\{t \in [0,1] : \sigma(t)/\mu(t) \le (2C_2)^{-1/2}\}$. For all $t \in (0, t_m]$ and $\boldsymbol{x}_t$,*

$$\tau^2 \nabla_{\boldsymbol{x}_t} \log p_t^{(\tau)}(\boldsymbol{x}_t) = \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) + \mathcal{E}_t(\boldsymbol{x}_t; \tau, p_0, \mu, \sigma), \qquad \|\mathcal{E}_t\| \le \frac{2C_3 d(\tau^2+1)}{\mu(t_m)^3} \sigma(t)^2. \tag{7}$$

uniformly in $\boldsymbol{x}_t$.

This result suggests two design implications. First, the $O(\sigma(t)^2)$ error rate motivates using low temperature only during the late, low-noise stage of sampling, where the score-reuse error is controlled.

Second, replacing the ideal tempered score $\tau^2 \nabla \log p_t^{(\tau)}$ by the pretrained score s_θ incurs both the temperature-reuse error bounded above and the network approximation error. This motivates the error analysis in the next section, which characterizes the instantaneous effect of stochasticity under score error.

Here $U = -\log p_0 + \text{const}$ is a statistical energy, not a bare molecular-mechanics potential. The global derivative bounds are a mathematical idealization: steric clashes and short-range repulsion can create singular or steep regions in physical energy landscapes. We interpret the theorem as characterizing locally regular, high-probability regions near the low-energy data manifold; a formal localization with explicit tail terms is left for future work.

2.3 Theoretical Analysis of Stochasticity

Although the upper-bound analysis of Xue et al. [2024a] recommends $\beta = 1$, larger stochasticity can empirically improve sample quality when sufficiently many sampling steps are used, for example in image generation (Sec. 6.3 of [Xue et al., 2024a]). We therefore analyze the exact continuous-time evolution of the discrepancy between (P2) and (P3). Define the total score error as $\delta_t(\mathbf{x}_t) \triangleq s_\theta(\mathbf{x}_t, t) - \tau^2 \nabla \log p_t^{(\tau)}(\mathbf{x}_t)$. It decomposes as

$$\delta_t = \underbrace{(s_\theta - \nabla \log p_t)}_{\text{network error}} + \underbrace{(\nabla \log p_t - \tau^2 \nabla \log p_t^{(\tau)})}_{\text{temperature error } (-\mathcal{E}_t \text{ from Theorem 2.1})} .$$

Theorem 2.2 (Evolution of Sampling Error)**.** *Assume that all involved distributions are smooth, strictly positive, and decay sufficiently fast so that the Fokker-Planck equation holds and all boundary terms vanish under integration by parts. Denote $\tilde{t} = T - t$ as the reverse time variable. The time derivative of the KL divergence satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{t}} D_{KL}(\tilde{p}_{\tilde{t}} \,\|\, p_{\tilde{t}}^{(\tau)}) = -\frac{\beta(\tilde{t})\tau^2 g_{\tilde{t}}^2}{2} \mathcal{F}(\tilde{p}_{\tilde{t}} \| p_{\tilde{t}}^{(\tau)}) + \frac{(1+\beta(\tilde{t}))g_{\tilde{t}}^2}{2} \int \tilde{p}_{\tilde{t}} \nabla \log \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} \cdot \delta_{\tilde{t}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{8}$$

where $\mathcal{F}(\tilde{p}_{\tilde{t}} \| p_{\tilde{t}}^{(\tau)}) = \int \tilde{p}_{\tilde{t}} \left\| \nabla \log(\tilde{p}_{\tilde{t}} / p_{\tilde{t}}^{(\tau)}) \right\|^2 \mathrm{d}\boldsymbol{x}$ is the relative Fisher information.

Theorem 2.2 reveals two opposing effects. The first term is non-positive and drives the sampler toward the target, with strength scaling in β. The second term, arising from score discrepancy $\delta_{\tilde{t}}$, can be either positive or negative. The KL divergence decreases strictly , at times with $\beta\tau^2 g^2 \mathcal{F} > 0$, when

$$\sup_{\boldsymbol{x}} \|\delta_{\tilde{t}}(\boldsymbol{x})\| < \frac{\beta(\tilde{t})\tau^2}{1+\beta(\tilde{t})} \sqrt{\mathcal{F}(\tilde{p}_{\tilde{t}} \| p_{\tilde{t}}^{(\tau)})}. \tag{9}$$

Under this condition and for fixed current distributions, larger β makes the instantaneous upper bound more negative and relaxes the admissible score-error threshold.

Overall, temperature annealing and tuned stochasticity enable effective sampling for sharply peaked molecular geometries. Full theorem proofs are deferred to Appendix B, and we derive the practical algorithm in the next section.

3 Temperature-Regulated Stochastic Sampling

3.1 Unified Discrete Formulation

To obtain implementable sampling updates, we discretize the reverse SDE and use a γ_t-parameterized conditional mean to express its drift. This framework relates the discrete-time parameter γ_t to the continuous-time stochasticity parameter $\beta(t)$, revealing how different choices of β affect inter-temporal correlations. Complete derivations are included in Appendix B.4.

We discretize the reverse SDE equation 1 using first-order Euler-Maruyama:

$$\boldsymbol{x}_{t-\Delta t} = \frac{\mu_{t-\Delta t}}{\mu_t} \boldsymbol{x}_t + \frac{(1+\beta(t))g_t^2 \Delta t}{2} s_\theta(\boldsymbol{x}_t, t) + \tau \sqrt{\beta(t) g_t^2 \Delta t}\, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

In this formulation, the update samples from a Gaussian distribution conditioned on the current state. The conditional expectation $\mathbb{E}[\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t]$ can be expressed analytically in terms of the score function.

Following Song et al. [2021a], we use the conditional mean of the following γ_t-parameterized family to reparameterize the drift:

$$p_{\gamma,\lambda}(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\Big(\mu_{t-\Delta t}\boldsymbol{x}_0 + \gamma_t \frac{\boldsymbol{x}_t - \mu_t \boldsymbol{x}_0}{\sigma_t},\ \tau^2 \lambda_t^2 \mathbf{I}\Big), \tag{11}$$

where $\gamma_t^2 + \lambda_t^2 = \sigma_{t-\Delta t}^2$ and $|\gamma_t| \leq \sigma_{t-\Delta t}$. The factor τ^2 in the covariance follows from the temperature-scaled forward noise in equation 4. Matching the $\mathbb{E}[\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t]$ term in equation 10 with this conditional distribution yields the γ_t-parameterized first-order discrete update:

$$\boldsymbol{x}_{t-\Delta t} = \frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + \Big(\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t\Big)\,\sigma_t\, s_\theta(\boldsymbol{x}_t, t) + \tau\sqrt{2\sigma_t(\sigma_{t-\Delta t} - \gamma_t)}\,\boldsymbol{\epsilon}. \tag{12}$$

The noise coefficient in this update is obtained using the first-order approximation $\beta(t)g_t^2\Delta t \approx 2\sigma_t(\sigma_{t-\Delta t} - \gamma_t)$. For bounded β, the update matches the leading-order SDE coefficients. In contrast, StoMax ($\gamma_t = 0$), introduced below, corresponds to $\beta_{\Delta t} = O(\Delta t^{-1})$ and has non-vanishing increments as $\Delta t \to 0$ at a fixed positive noise level. The update's noise variance generally differs from the conditional-family variance $\tau^2(\sigma_{t-\Delta t}^2 - \gamma_t^2)$ at finite step size; for StoMax, this discrepancy persists as $\Delta t \to 0$. The correspondence between the continuous-time parameter $\beta(t)$ and the discrete-time parameter γ_t is summarized in Table 1.

3.2 Representative Variants

Under the unified formulation above, we reexamine commonly used sampling strategies through the lens of inter-temporal correlation structure. (1) The Markov coupling of the associated DDIM family, i.e. $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0) = p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t})$, gives $\gamma_t = \frac{\mu_t \sigma_{t-\Delta t}^2}{\mu_{t-\Delta t}\sigma_t}$. Propositions B.4 and B.5 show that this coefficient recovers the usual DDPM and BFN parameterization; in our first-order update, it corresponds to $\beta(t) \to 1$ as $\Delta t \to 0$. (2) The deterministic limit $\beta = 0$ gives $\gamma_t = \sigma_{t-\Delta t}$, aligning with the DDIM formulation.

These two strategies provide deterministic and standard-stochasticity references. A natural question arises: Can intermediate or extreme choices yield better performance? To investigate this, we consider an underexplored variant: (3) Setting $\gamma_t = 0$ gives a natural high-stochasticity reference. In the associated conditional family equation 11, this choice yields $p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_0)$, so the noise is fully refreshed conditional on $\boldsymbol{x}_0$. It also maximizes the conditional variance $\tau^2(\sigma_{t-\Delta t}^2 - \gamma_t^2)$ over the full admissible range $|\gamma_t| \leq \sigma_{t-\Delta t}$. In the practical update equation 12, $\gamma_t = 0$ maximizes the noise coefficient within $0 \leq \gamma_t \leq \sigma_{t-\Delta t}$. We call this variant Stochasticity Maximized Sampling (StoMax). Negative γ_t corresponds to anticorrelated conditional couplings and can further increase the practical noise coefficient. A systematic study of this regime is left for future work.

Table 1: Correspondence between $\beta(t)$ and the discrete-time parameter γ_t as $\Delta t \to 0$.

$\beta(t)$	γ_t	**Correlation Assumption**	**Representative Regime**
0	$\sigma_{t-\Delta t}$	$p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ is a point mass	DDIM/deterministic sampling
1	$\frac{\mu_t \sigma_{t-\Delta t}^2}{\mu_{t-\Delta t}\sigma_t}$	$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0) = p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t})$	DDPM/BFN sampling
$\to \infty$	0	$p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_0)$	StoMax sampling

3.3 Temperature Annealing Strategy

Theorem 2.1 gives a finite window $0 < t \leq t_m$ with $O(\sigma^2)$ error, motivating temperature reduction only during the late, low-noise stage of sampling. Since the theorem is stated for fixed τ, it provides a pointwise-in-time error bound that guides the design of $\tau(t)$, though it does not guarantee that the time-varying annealed process has an exact tempered marginal. Prior work has shown that naïvely scaling the Gaussian noise by a constant temperature factor often produces over-smoothed samples in image generation [Dhariwal and Nichol, 2021]. We observe a similar limitation: constant low-temperature sampling fails to adequately explore distant high-density regions.

Table 2: Comparison of vanilla and TREASSURE(γ, ν) sampling on QM9 unconditional generation.

Models(Sampling)	Atom sta. (%) ↑	Mol sta. (%) ↑	Valid (%) ↑	V×U (%)↑
Data	99.0	95.2	97.7	97.7
EDM(Vanilla)	98.7	82.0	91.9	90.7
EDM$(0, 0)$	98.9	87.9	94.5	92.1
EDM$(0, 0.3)$	**99.2**	**92.0**	**96.5**	**92.8**
UniGEM$_{\text{EDM}}$(Vanilla)	98.9	89.2	94.7	92.8
UniGEM$_{\text{EDM}}$($\gamma_m, 0.3$)	99.1	92.0	96.1	92.4
UniGEM$_{\text{EDM}}$$(0, 0)$	**99.6**	**95.6**	**97.8**	**95.9**
GeoBFN(Vanilla)	99.3	93.0	96.5	**92.7**
GeoBFN$(0, 0)$	99.1	92.1	95.2	91.0
GeoBFN$(0, 0.5)$	**99.3**	**94.2**	**96.9**	91.9
UniGEM$_{\text{BFN}}$(Vanilla)	99.3	93.7	97.3	**93.0**
UniGEM$_{\text{BFN}}$$(0, 0)$	99.3	94.2	96.7	92.3
UniGEM$_{\text{BFN}}$$(0, 0.5)$	**99.5**	**95.7**	**97.8**	91.3

Algorithm 1 Temperature-Annealed Stochastic Sampling (TREASSURE)

Require: Trained score network s_θ with noise schedule $\mu(t), \sigma(t)$; initial distribution $p_T(\boldsymbol{x})$; number of steps T; annealing exponent ν; mean parameter γ_t

1: $\boldsymbol{x} \sim p_T(\boldsymbol{x})$
2: **for** $i = T$ **downto** 1 **do**
3: $\quad \tau \leftarrow (i/T)^\nu$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4: $\quad \boldsymbol{x} \leftarrow \left(\frac{\mu_{i-1}}{\mu_i}\sigma_i - \gamma_i\right)\sigma_i s_\theta(\boldsymbol{x}, i) + \frac{\mu_{i-1}}{\mu_i}\boldsymbol{x} + \tau\sqrt{2\sigma_i(\sigma_{i-1} - \gamma_i)}\,\boldsymbol{\epsilon}.$
5: **end for**
6: **return** $\boldsymbol{x}$

To mitigate both concerns, we adopt a time-annealing temperature scheme: The temperature starts at one and decreases along the reverse trajectory. Smaller exponents retain higher temperatures for longer exploration, whereas larger exponents induce earlier and more aggressive cooling to quickly concentrate samples within low-energy regions. Concretely, we use the power-law schedule $\tau(t) = (t/T)^\nu, \quad t \in [0, T]$, where ν governs the temperature decay rate. Empirically, this power-law annealing schedule performs strongly among the competing designs considered, with results provided in Appendix C.3. Our full sampling algorithm is summarized in Algorithm 1.

4 Experiments

We benchmark TREASSURE ranging from small molecules to protein complexes and examine the empirical implications of the two parameters. Throughout, "sample quality" denotes chemical plausibility, proximity to the molecular manifold, and adherence to requested conditions. It is measured empirically by task-specific metrics, stability/validity, lDDT/DockQ, rotamer quality/pLDDT, not by closeness to empirical training distribution. The temperature-induced reweighting is purely distributional and does not access or optimize any evaluation metric. Diversity and coverage are reported separately.

4.1 Small Molecule Generation

Unconditional molecular generation evaluates a generative model's ability to capture the overall data distribution, aiming to produce molecules that are both chemically valid and structurally diverse. We benchmark our method on two widely used datasets, following the dataset splits and evaluation protocols of prior work [Hoogeboom et al., 2022, Song et al., 2024a, Feng et al., 2025]. Specifically, we generate 10,000 molecules per model with 1000 sampling steps and predict bond types based on inter-atomic distances as in Hoogeboom et al. [2022].

4.1.1 Results on QM9

The QM9 dataset [Ruddigkeit et al., 2012, Ramakrishnan et al., 2014] contains approximately 134,000 small organic molecules with up to nine heavy atoms and serves as a standard benchmark for generative models. We compare our approach with four diffusion-based baselines. Equivariant

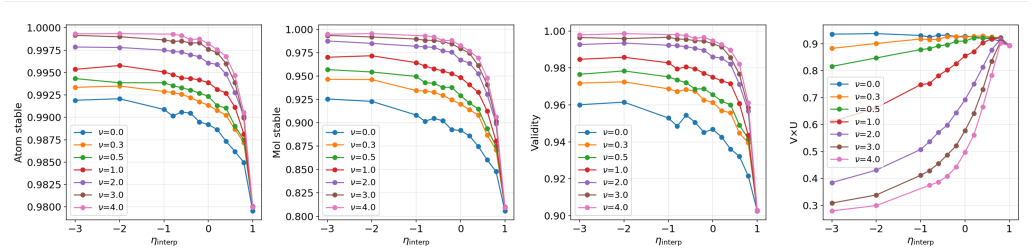


Figure 2: Joint variation of temperature and stochasticity by interpolating from DDIM to DDPM.

Diffusion Model (EDM) [Hoogeboom et al., 2022] and its UniGEM [Feng et al., 2025] variant adopt the DDPM [Ho et al., 2020], while GeoBFN[Song et al., 2024a] and its UniGEM variant are built upon the Bayesian Flow Networks [Graves et al., 2023]. EDM and GeoBFN jointly diffuse atom positions and atom types, whereas their UniGEM counterparts only diffuse atom positions and predict atom types subsequently. Here and throughout, "vanilla" denotes each released model's original/default sampler. For the small-molecule models, our Markov-type reference uses $\nu = 0$ and $\gamma = \gamma_m$ in equation 12; Appendix B.5 relates this coefficient to the usual DDPM/BFN parameterization in the small-step limit. For protein models, vanilla corresponds to ODE-Churn sampling for protein modeling, with detail described in F. Results are summarized in Table 2.

In contrast, our method employs TREASSURE sampling with tunable mean parameter γ and temperature annealing exponent ν. Across all models, selected configurations improve stability and validity, while V×U can decrease as probability mass becomes more concentrated. We analyze this trade-off via manifold coverage in Appendix C.4, which shows improved precision and test coverage with a modest recall reduction.

4.1.2 Results on GEOM-Drugs

We further validate TREASSURE on the larger, structurally more complex GEOM-Drugs dataset, which includes 430,000 molecules with up to 181 atoms. Table 3 compares different sampling configurations on EDM and UniGEM-EDM. We follow prior work [Hoogeboom et al., 2022] and report atomic stability and validity as quality metrics. The conclusions align with QM9: higher stochasticity and lower temperature can yield superior generation performance.

Table 3: Comparison of vanilla and TREASSURE(γ, ν) sampling on GEOM-Drugs unconditional generation.

Models(Sampling)	Atom sta. (%)↑	Valid (%) ↑
Data	86.5	99.9
EDM(Vanilla)	81.3	92.6
EDM($\gamma_m, 3$)	**91.3**	93.0
EDM$(0, 0)$	86.2	**99.7**
$\text{UniGEM}_{\text{EDM}}$(Vanilla)	85.1	98.5
$\text{UniGEM}_{\text{EDM}}(\gamma_m, 1)$	**92.2**	95.8
$\text{UniGEM}_{\text{EDM}}(0, 0)$	89.5	**99.9**

4.1.3 Design Space Exploration

To study the joint design space, we interpolate/extrapolate stochasticity from DDIM to DDPM and vary the temperature exponent $\nu \in [0, 4]$ on UniGEM(EDM):

$$\gamma_t = \eta\gamma_t^{\text{DDIM}} + (1-\eta)\gamma_t^{\text{DDPM}} = \eta\sigma_{t-\Delta t} + (1-\eta)\frac{\mu_t\sigma_{t-\Delta t}^2}{\mu_{t-\Delta t}\sigma_t}, \quad \eta \in (-\infty, 1]. \tag{13}$$

Results are shown in Fig. 2.

Over a broad region, increasing sampling stochasticity and the temperature-annealing exponent improves molecular stability and validity. By contrast, uniqueness declines under extreme stochasticity and temperature configurations, revealing an inherent stability–diversity tradeoff.

4.2 Protein Structure Prediction

To assess the generalizability of TREASSURE beyond small molecules, we apply it to protein structure prediction using RoseTTAFold3 (RF3) [Corley et al., 2025], a leading model for biomolecular complex prediction. RF3 generates structures via a diffusion model whose training and sampling

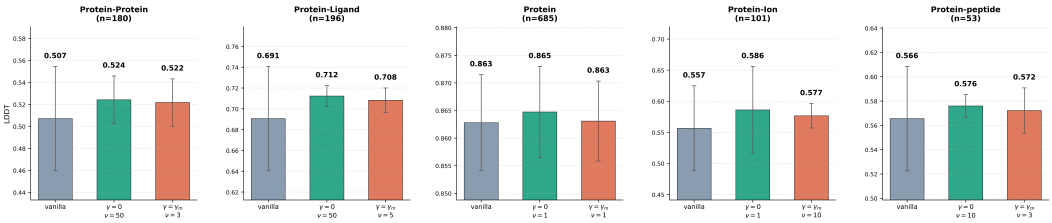


Figure 3: Comparison of LDDT scores across validation subsets using various sampling methods in RF3. Error bars denote standard deviations across samples with different diffusion seeds.

algorithms are inspired by [Karras et al., 2022], aligning closely with prominent structure prediction models such as AlphaFold3. In its default setting, RF3 samples structures using an ODE integrator with stochastic "churn" operations that inject noise during the denoising process. We replace the vanilla sampler with TREASSURE to systematically control the sharpness of the sampling distribution and improve sample quality in a single draw. We evaluate two representative configurations: the Markov variant $\gamma = \gamma_m$ and StoMax variant $\gamma = 0$, each combined with different temperature annealing exponents ν. We evaluate on protein monomers and protein complexes from the recent PDB test set following the RF3 evaluation protocol.

Results are shown in Figure 3. All methods use paired random seeds, identical MSAs, and five samples per diffusion batch. We observe higher mean lDDT with smaller empirical standard deviations for some subsets. This aligns with our core motivation of steering samples toward sharper, higher-probability data distributions. We also evaluate the antibody-antigen benchmark from AlphaFold3 [Abramson et al., 2024]. Results in Figure 5 show the same trend: on the same 224 paired interfaces, our methods raise DockQ success (> 0.23) from 37.9% to 42.0%.

4.3 De Novo Protein Design

We further evaluate TREASSURE on protein backbone generation using RFdiffusion3, a diffusion model that generates protein structures [Butcher et al., 2025]. RFdiffusion3 is an all-atom biomolecular diffusion model with an AlphaFold3-inspired transformer denoising architecture, trained primarily on Alphafold Database [Varadi et al., 2022] and experimentally deposited PDB structures [Berman et al., 2000]. Its sampling algorithm, like AlphaFold3 and RF3, employs a churn ODE sampler. We replaced the baseline sampler with TREASSURE and evaluated their performance across 100 unconditionally generated structures by RFdiffusion3, with lengths ranging from 100 to 300 residues.

We evaluate the structures based on designability, side-chain rotamer quality, pLDDT and structural diversity, with detailed descriptions in D. Among all metrics, rotamer quality better reflects physical validity which measures consistency with canonical rotamer distributions. Besides, de novo protein design requires a balance between quality and diversity for broad application scenarios. To systematically compare methods, we partition all results into four diversity intervals and compare quality metrics within each regime. We also tune two adjustable parameters of vanilla sampling, noise scale g and step size s, to cover a wide range of quality-diversity trade-offs for fair comparison.

Results are shown in Table 4. TREASSURE achieves consistently lower rotamer loss across the same diversity interval. This advantage indicates improved physical plausibility of generated structures. At comparable diversity, TREASSURE$(\gamma_m, 0.5)$ improves all three quality metrics and constitutes a Pareto point rather than strictly dominating every diversity metric. Moreover, by increasing the temperature annealing exponent ν and stochasticity, TREASSURE can explore higher-quality regions with reduced diversity. These results demonstrate that TREASSURE can regulate the quality-diversity trade-off in a principled and controllable manner.

4.4 Supplementary Validation and Ablation Study

We summarize key findings in this section, with detailed experimental results deferred to Appendix C. *Firstly*, the synthetic studies exhibit the non-trivial validity window and late-stage decay of score-reuse error predicted by Theorem 2.1. The molecular analyses further show that temperature annealing sharpens bond-length distributions and reduces their standard deviation by over 30%, consistent with the intended sharpness control of τ. *Secondly*, increased stochasticity is empirically more robust to

Table 4: Experimental results on RFdiffusion3. Vanilla sampler parameters: g = noise scale, s = step size. TREASSURE parameters: γ = mean parameter, ν = temperature annealing exponent.

Sampling Method	Rotamer Loss↓	pLDDT↑	Designability↑	Diversity(TM Score)↓	Diversity(AUC)↑
Vanilla (g=0.2,s=1.0)	0.40	70.48	0.45	0.36	0.55
Vanilla (g=0.6,s=1.0)	0.40	74.73	0.53	0.38	0.52
TREASSURE($\gamma_m, 0.5$)	**0.26**	**76.22**	**0.59**	0.37	0.52
Vanilla (g=0.8,s=1.0)	0.40	76.59	0.63	0.39	0.50
Vanilla (g=0.8,s=1.5)	0.30	82.75	0.82	0.41	0.47
Vanilla (g=0.2,s=1.5)	0.30	84.32	0.81	0.41	0.46
Vanilla (g=0.6,s=1.5)	0.32	**85.18**	**0.87**	0.43	0.42
TREASSURE($\gamma_m, 3$)	**0.22**	82.26	0.76	0.42	0.43
TREASSURE$(0, 1)$	0.25	83.04	0.74	0.43	0.39
TREASSURE($\gamma_m, 5$)	**0.22**	**86.15**	**0.86**	0.45	0.37
TREASSURE($\gamma_m, 10$)	0.24	85.24	0.68	0.54	0.32
TREASSURE($\gamma_m, 50$)	**0.24**	84.84	0.75	0.54	0.30
TREASSURE$(0, 5)$	0.31	**89.52**	**0.85**	0.53	0.24
TREASSURE$(0, 50)$	0.32	87.54	0.71	0.56	0.23

initialization and step reduction: StoMax uses $10\times$ fewer denoising steps than the 1000-step DDPM reference while attaining higher U$\times$V. Although Theorem 2.2 is continuous-time and does not model discretization error, these finite-step results suggest that its error-correction mechanism can remain robust in the presence of such error. *Thirdly*, on design choices, we systematically compare the effects of different temperature schedule curves on generative performance, and find the power-law schedule delivers strong empirical performance. Limitations of the proposed approach are discussed in Section G.

5 Conclusion

We present TREASSURE, a training-free framework controlling target sharpness (τ) and stochastic correction (β) in diffusion-based molecular generation. Our theory gives an explicit score-reuse window under bounded second- and third-order derivatives, without a Gaussian parametric assumption, together with an exact KL-evolution identity that characterizes instantaneous stochastic correction. Empirically, TREASSURE consistently improves physical validity and structural quality across small-molecule generation, protein structure prediction, and de novo protein design, without retraining or any auxiliary model. It can be plugged directly into existing pretrained diffusion models, offering a simple yet principled recipe for sharper, higher-quality molecular generation.

Acknowledgements

This work is supported by Innovative Drug Research and Development-National Science and Technology Major Project (No.2025ZD1802501) and Beijing Frontier Research Center for Biological Structure Fundings. The authors declare no competing interests.

References

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

Shikun Feng, Yuyan Ni, Lu yan, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. UniGEM: A unified approach to generation and property prediction for molecules. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=Lb91pXwZMR`.

Yuxuan Song, Jingjing Gong, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified generative modeling of 3d molecules with bayesian flow networks. In *The Twelfth International Conference on Learning Representations*, 2024a.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Nathaniel Corley, Simon Mathis, Rohith Krishna, Magnus S Bauer, Tuscan R Thompson, Woody Ahern, Maxwell W Kazman, Rafael I Brent, Kieran Didi, Andrew Kubaney, et al. Accelerating biomolecular modeling with atomworks and rf3. *BioRxiv*, 2025.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I Brent, Yanjing Li, Nathaniel Corley, Paul T Kim, Jonathan Funk, Simon Mathis, Saman Salike, et al. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, 2025.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2021a. URL `https://openreview.net/forum?id=St1giarCHLP`.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023.

Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 560–576. Springer, 2023.

Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks. *arXiv preprint arXiv:2308.07037*, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

Shuchen Xue, Mingyang Yi, Weijian Luo, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhi-Ming Ma. Sa-solver: Stochastic adams solver for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024a.

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

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Ludwig Boltzmann. Studien uber das gleichgewicht der lebenden kraft. *Wissenschafiliche Abhandlungen*, 1:49–96, 1868.

Yuyan Ni, Shikun Feng, Haohan Chi, Bowen Zheng, Huan ang Gao, Wei-Ying Ma, Zhi-Ming Ma, and Yanyan Lan. Straight-line diffusion model for efficient 3d molecular generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Kaiwen Xue, Yuhao Zhou, Shen Nie, Xu Min, Xiaolu Zhang, JUN ZHOU, and Chongxuan Li. Unifying bayesian flow networks and diffusion models through stochastic differential equations. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=1jHiq640y1.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0234c510bc6d908b28c70ff313743079-Paper.pdf.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

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

Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie 2. *arXiv preprint arXiv:2405.15489*, 2024.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. In *International Conference on Learning Representations (ICLR)*, 2025a.

Yanbo Xu, Yu Wu, Sungjae Park, Zhizhuo Zhou, and Shubham Tulsiani. Temporal score rescaling for temperature sampling in diffusion and flow models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=WbfRe0cheB.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alan Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Vhc0KrcqWu.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-proteina: Atomistic protein generation via partially latent flow matching. *arXiv preprint arXiv:2507.09466*, 2025b.

Michael Samuel Albergo, Mark Goldstein, Nicholas Matthew Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 921–937. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/albergo24a.html`.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36, 2024b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Efficient 3d molecular generation with flow matching and scale optimal transport. In *ICML 2024 AI for Science Workshop*, 2024.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pages 2024–11, 2025.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, et al. Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. *BioRxiv*, pages 2025–01, 2025.

Chengyue Gong, Xinshi Chen, Yuxuan Zhang, Yuxuan Song, Hao Zhou, and Wenzhi Xiao. Protenix-mini: Efficient structure predictor via compact architecture, few-step diffusion and switchable plm. *arXiv preprint arXiv:2507.11839*, 2025.

Bo Qiang, Chengyue Gong, Xinshi Chen, Yuxuan Zhang, and Wenzhi Xiao. Protenix-mini+: efficient structure prediction model with scalable pairformer. *arXiv preprint arXiv:2510.12842*, 2025.

Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, et al. Atom-level enzyme active site scaffolding using rfdiffusion2. *Nature Methods*, 23(1):96–105, 2026.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

A Motivation

Under the Boltzmann distribution [Boltzmann, 1868], molecular distributions are inherently governed by the potential energy landscape. Small deviations in covalent bond lengths or bond angles induce a sharp rise in energy, confining valid molecular configurations to narrow low-energy regions, as illustrated in Figure 4. This intrinsic concentration stands in stark contrast to natural images, implying generative methods optimized for image modeling cannot be naively transferred to molecular modeling.

In particular, we observe that molecules generated by standard diffusion samplers can exhibit noticeably flatter bond-length and bond-angle distributions such as in Figure 4, with a non-negligible fraction of samples falling into high-energy or physically invalid regions. This discrepancy motivates us to develop sampling strategies that better align the generated distribution with the intrinsically sharp molecular distribution.

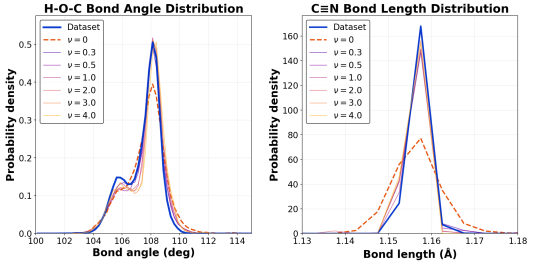


Figure 4: Vanilla DDPM ($\nu = 0$) smooths the H-O-C angle and C≡N length distributions, while increasing ν sharpens them toward the data modes.

B Theoretical Proofs

B.1 Temperature Scaling for Langevin Dynamics

For Langevin dynamics, the sampling temperature can be controlled by adding a constant τ in front of the stochastic term:

$$\mathrm{d}\boldsymbol{x} = \frac{1}{2} g(t)^2 \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\, \mathrm{d}t + \tau g(t)\, \mathrm{d}w'. \tag{14}$$

where τ is the temperature parameter. It can be shown that $\pi(\boldsymbol{x}) \propto p(\boldsymbol{x})^{\frac{1}{\tau^2}}$ is the stationary distribution of the above process:

Proof. The marginal density $\pi_t(x)$ satisfies the Fokker-Planck equation:

$$\frac{\partial \pi_t(\boldsymbol{x})}{\partial t} = -\nabla \cdot \left[\frac{1}{2} g(t)^2 \nabla_{\boldsymbol{x}} \log p(x) \pi_t(\boldsymbol{x})\right] + \frac{1}{2} \nabla \cdot \nabla \cdot \left[\tau^2 g(t)^2 \pi_t(\boldsymbol{x})\right]. \tag{15}$$

At stationarity, the density no longer changes with time, i.e., $\frac{\partial \pi_t(x)}{\partial t} = 0$, so we have

$$\nabla \cdot \left[\frac{1}{2} g(t)^2 \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) \pi(\boldsymbol{x})\right] = \frac{1}{2} \nabla \cdot \nabla \cdot \left[\tau^2 g(t)^2 \pi(\boldsymbol{x})\right]. \tag{16}$$

One can verify that $\pi(\boldsymbol{x}) \propto p(\boldsymbol{x})^{\frac{1}{\tau^2}}$ satisfies the above equation. □

As τ increases, p^{1/τ^2} becomes flatter whenever it remains normalizable. As $\tau \to 0$, its mass concentrates on the set of global maximizers of p; a single Dirac limit additionally requires a unique isolated maximizer.

However, temperature scaling methods for diffusion model sampling have been less studied. [Ingraham et al., 2023] proposed weighting the score function, while [Dhariwal and Nichol, 2021] suggested weighting the Gaussian noise in sampling. The derivation of the former assumes the data follows an isotropic Gaussian distribution, an assumption far from real-world data, while the

latter lacks rigorous proof. Without a Gaussian parametric assumption, we show under the regularity conditions of Theorem 2.1 that Gaussian-noise scaling approximates temperature scaling in the low-noise regime.

B.2 Temperature Scaling for Diffusion Models

This section shows that temperature-scaled diffusion sampling approximates a target $p_0^{(\tau)} \propto \exp(-U/\tau^2)$ using a forward process with noise scale $\tau\sigma(t)$. In the late reverse stage, the quadratic term dominates and the score-reuse error becomes small, motivating annealed sampling.

Interpretation of Temperature Scaling Let the data distribution be $p_0(\boldsymbol{x}_0) \propto e^{-U(\boldsymbol{x}_0)}$, where U is a statistical energy. Consider the forward diffusion process (t from 0 to 1):

$$\boldsymbol{x}_t = \mu(t)\boldsymbol{x}_0 + \sigma(t)\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \tag{17}$$

where $\mu(t)$ and $\sigma(t)$ satisfy the initial conditions $\mu(0) = 1$, $\sigma(0) = 0$, as well as $\mu(t) \geq 0$ non-increasing, $\sigma(t) \geq 0$, and $\frac{\mathrm{d}(\sigma(t)/\mu(t))}{\mathrm{d}t} \geq 0$ (proof provided in Appendix A.1 in [Ni et al., 2025]). The marginal distribution is then

$$p(\boldsymbol{x}_t) = \int \frac{1}{Z} e^{-U(\boldsymbol{x}_0)} \mathcal{N}(\boldsymbol{x}_t; \mu(t)\boldsymbol{x}_0, \sigma(t)^2 I)\, d\boldsymbol{x}_0. \tag{18}$$

Taking the logarithm and ignoring constants independent of $\boldsymbol{x}_t$:

$$\log p(\boldsymbol{x}_t) = \log \int \exp\left(-U(\boldsymbol{x}_0) - \frac{\|\boldsymbol{x}_t - \mu(t)\boldsymbol{x}_0\|^2}{2\sigma(t)^2}\right) d\boldsymbol{x}_0. \tag{19}$$

Define

$$f_{\boldsymbol{x}_t}(\boldsymbol{x}_0) = U(\boldsymbol{x}_0) + \frac{\|\boldsymbol{x}_t - \mu(t)\boldsymbol{x}_0\|^2}{2\sigma(t)^2}. \tag{20}$$

As $t \to 0$, we have $\sigma(t) \to 0$, so the quadratic term dominates. In this regime, $f_{\boldsymbol{x}_t}$ has a unique global minimum approximately at $\boldsymbol{x}_0^* \approx \boldsymbol{x}_t/\mu(t)$. Conversely, as $t \to 1$, $\mu(t)/\sigma(t) \to 0$, making the quadratic term negligible; then $f_{\boldsymbol{x}_t}(\boldsymbol{x}_0) \approx U(\boldsymbol{x}_0)$, and its landscape is governed primarily by U, potentially exhibiting multiple minima of comparable magnitude.

Now consider the temperature-scaled distribution $p_0^{(\tau)} \propto \exp(-U/\tau^2)$ and the modified forward process:

$$\boldsymbol{x}_t = \mu(t)\boldsymbol{x}_0 + \tau\sigma(t)\boldsymbol{\epsilon}, \tag{21}$$

i.e., keeping the mean coefficient unchanged and scaling the variance by τ^2. Then we have

$$\begin{aligned} \log p^{(\tau)}(\boldsymbol{x}_t) &= \log \int \exp\left(-\frac{1}{\tau^2}\, U(\boldsymbol{x}_0) - \frac{\|\boldsymbol{x}_t - \mu(t)\boldsymbol{x}_0\|^2}{2\sigma(t)^2\tau^2}\right) d\boldsymbol{x}_0 \\ &= \log \int \exp\left(-\frac{1}{\tau^2} f_{\boldsymbol{x}_t}(\boldsymbol{x}_0)\right) d\boldsymbol{x}_0. \end{aligned} \tag{22}$$

Compared to equation 19, the expression inside the integral is scaled by $1/\tau^2$.

Heuristically, as $t \to 0$ we have $\sigma(t) \to 0$, so the Hessian of $f_{\boldsymbol{x}_t}$ scales like $\mu(t)^2/\sigma(t)^2 \to \infty$ and the log integrand $e^{-f_{\boldsymbol{x}_t}/\tau^2}$ concentrates ever more sharply around the *same* τ-independent minimizer $\boldsymbol{x}_0^*$. Consequently both $\log p^{(\tau)}(\boldsymbol{x}_t)$ and $\log p(\boldsymbol{x}_t)$ are governed by a Laplace expansion about $\boldsymbol{x}_0^*$ whose leading terms differ only through the global prefactor $1/\tau^2$; hence $\nabla_{\boldsymbol{x}_t} \log p^{(\tau)}(\boldsymbol{x}_t) \approx \frac{1}{\tau^2}\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$. This concentration is driven by $\sigma \to 0$ at fixed τ, not by $\tau \to 0$. The error of this approximation is quantified rigorously in the next paragraph.

To sample from the temperature-scaled distribution, we need to use the reverse process with respect to $p^{(\tau)}(\boldsymbol{x}_t)$ and using scaled noise schedule $(\mu(t), \sigma(t)\tau)$. As shown in [Xue et al., 2024a, Karras et al., 2022], the backward SDE that shares the same Fokker-Planck equation as the forward process equation 5 is given by:

$$d\boldsymbol{x}_t = \left[\frac{\dot{\mu}_t}{\mu_t}\boldsymbol{x}_t - \frac{1+\beta(t)}{2}\tau^2 g_t^2 \nabla_{\boldsymbol{x}_t} \log p^{(\tau)}(\boldsymbol{x}_t)\right] \mathrm{d}t + \sqrt{\beta(t)}\tau g_t \mathrm{d}w_t, \tag{23}$$

Substituting the approximations $\nabla_{\boldsymbol{x}_t} \log p^{(\tau)}(\boldsymbol{x}_t) \approx \frac{1}{\tau^2} \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ and $s_\theta \approx \nabla_{\boldsymbol{x}_t} \log p$ into equation 23, we obtain equation 1. This shows that, up to the approximations of score function, simply multiplying the diffusion term by the temperature τ controls the temperature of the sampling distribution.

Approximation Error Analysis

Theorem B.1 (Temperature Scaling Approximation)**.** *Assume that*

1. $p_0(\boldsymbol{x}) = Z^{-1} e^{-U(\boldsymbol{x})}$ *and* $p_0^{(\tau)}(\boldsymbol{x}) = Z_\tau^{-1} e^{-U(\boldsymbol{x})/\tau^2}$ *are normalizable,* $U \in C^3(\mathbb{R}^d)$*, and* $\|\nabla^2 U(\boldsymbol{x})\|_{\mathrm{op}} \le C_2$*,* $\|\nabla^3 U(\boldsymbol{x})\|_{\mathrm{op}} \le C_3$ *for all* $\boldsymbol{x}$*;*
2. *The noise schedule* $\mu(t)$ *and* $\sigma(t)$ *are continuous on* $[0,1]$ *and positive on* $(0,1)$*,* $\mu(0) = 1$*,* $\sigma(0) = 0$*, with* $\mu(t)$ *non-increasing and* $\sigma(t)/\mu(t)$ *non-decreasing on* $[0,1]$ *(as in Sec. B.2);*
3. $\tau > 0$ *is fixed and independent of* t*.*

For $C_2 > 0$*, define the strong-convexity threshold by*

$$t_m = \sup\left\{ t \in [0,1] : \frac{\sigma(t)}{\mu(t)} \le \frac{1}{\sqrt{2C_2}} \right\}, \tag{24}$$

and set $t_m = 1$ *if* $C_2 = 0$ *or if the inequality holds throughout* $[0,1]$*. Continuity and monotonicity of* σ/μ *imply that for* ***every*** $0 < t \le t_m$ *the map* $\boldsymbol{x}_0 \mapsto f_{\boldsymbol{x}_t}(\boldsymbol{x}_0)$ *is strongly convex with a unique minimizer* $\boldsymbol{x}_0^* = \boldsymbol{x}_0^*(\boldsymbol{x}_t, t)$*, and the temperature-scaling identity*

$$\tau^2 \, \nabla_{\boldsymbol{x}_t} \log p^{(\tau)}(\boldsymbol{x}_t) = \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t) + \mathcal{E}_t \tag{25}$$

holds with a non-asymptotic, uniform *error bound*

$$\boxed{\|\mathcal{E}_t\| \;\le\; M\,\sigma(t)^2 \quad \textit{for all } 0 < t \le t_m, \qquad M = \frac{2\,C_3\,d\,(\tau^2+1)}{\mu(t_m)^3}} \tag{26}$$

Proof. Our goal is to estimate the vector-valued error

$$\mathcal{E}_t \triangleq \tau^2 \nabla_{\mathbf{x}_t} \log p^{(\tau)}(\mathbf{x}_t) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t).$$

Differentiating $\log p^{(\tau)}(\boldsymbol{x}_t) = \log \int e^{-f_{\boldsymbol{x}_t}/\tau^2} \, d\boldsymbol{x}_0$ under the integral sign (justified by normalizability and the Gaussian factor) and using $\nabla_{\boldsymbol{x}_t} f_{\boldsymbol{x}_t}(\boldsymbol{x}_0) = \frac{\boldsymbol{x}_t - \mu \boldsymbol{x}_0}{\sigma^2}$,

$$\tau^2 \, \nabla_{\boldsymbol{x}_t} \log p^{(\tau)}(\boldsymbol{x}_t) = -\left\langle \nabla_{\boldsymbol{x}_t} f_{\boldsymbol{x}_t} \right\rangle_\tau = \frac{\mu \, \langle \boldsymbol{x}_0 \rangle_\tau - \boldsymbol{x}_t}{\sigma^2},$$

Here $\langle \cdot \rangle_a$ denotes expectation under the tempered posterior of $\boldsymbol{x}_0$ given $\boldsymbol{x}_t$ at temperature a:

$$\langle h(\boldsymbol{x}_0) \rangle_a := \int_{\mathbb{R}^d} h(\boldsymbol{x}_0) \, q_{a,t}(\boldsymbol{x}_0 \mid \boldsymbol{x}_t) \, d\boldsymbol{x}_0,$$

where, for any $a > 0$,

$$q_{a,t}(\boldsymbol{x}_0 \mid \boldsymbol{x}_t) := \frac{\exp\!\left[-f_{\boldsymbol{x}_t}(\boldsymbol{x}_0)/a^2\right]}{\int_{\mathbb{R}^d} \exp[-f_{\boldsymbol{x}_t}(\boldsymbol{z})/a^2] \; d\boldsymbol{z}}. \tag{27}$$

Setting $\tau = 1$ gives $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t) = \big(\mu \langle \boldsymbol{x}_0 \rangle_1 - \boldsymbol{x}_t\big)/\sigma^2$. Subtracting, the $\boldsymbol{x}_t/\sigma^2$ terms cancel and

$$\mathcal{E}_t = \frac{\mu}{\sigma^2} \Big(\langle \boldsymbol{x}_0 \rangle_\tau - \langle \boldsymbol{x}_0 \rangle_1 \Big).$$

Writing $I_n^{(\tau)} = \int e^{-f_{\boldsymbol{x}_t}/\tau^2} \boldsymbol{x}_0 \, d\boldsymbol{x}_0$ and $I_d^{(\tau)} = \int e^{-f_{\boldsymbol{x}_t}/\tau^2} d\boldsymbol{x}_0$ (so $\langle \boldsymbol{x}_0 \rangle_\tau = I_n^{(\tau)}/I_d^{(\tau)}$), the error is the difference of two *posterior means* of $\boldsymbol{x}_0$ at inverse temperatures $1/\tau^2$ and 1.

When $f_{\mathbf{x}_t}$ is purely quadratic, each posterior mean equals the unique minimizer of $f_{\mathbf{x}_t}$ independently of τ, so $\langle \boldsymbol{x}_0 \rangle_\tau = \langle \boldsymbol{x}_0 \rangle_1 = \boldsymbol{x}_0^*$ and $\mathcal{E}_t = \mathbf{0}$; the error is thus entirely due to the anharmonic (cubic and higher) part of U. As $t \to 0$ the quadratic term dominates and the two posteriors concentrate near $\boldsymbol{x}_0^*$; we now quantify the residual in terms of $\mu(t), \sigma(t)$.

Lemma B.2 (Strong convexity for $0 < t \le t_m$)**.** *For every $0 < t \le t_m$ (with t_m from equation 24) the function $\mathbf{x}_0 \mapsto f_{\mathbf{x}_t}(\mathbf{x}_0)$ is m_t-strongly convex on $\mathbb{R}^d$ with*

$$m_t \;=\; \frac{\mu(t)^2}{\sigma(t)^2} - C_2 \;\ge\; \frac{\mu(t)^2}{2\sigma(t)^2} \;> 0,$$

so it has a unique global minimizer $\boldsymbol{x}_0^*$, and $H \triangleq \nabla^2 f_{\boldsymbol{x}_t}(\boldsymbol{x}_0^*) \succeq m_t I$.

Proof. The Hessian is $\nabla^2 f_{\mathbf{x}_t}(\mathbf{x}_0) = \nabla^2 U(\mathbf{x}_0) + \frac{\mu^2}{\sigma^2} I$. The uniform bound $\|\nabla^2 U\| \le C_2$ gives $\nabla^2 U(\boldsymbol{x}_0) \succeq -C_2 I$, hence $\nabla^2 f_{\boldsymbol{x}_t} \succeq (\frac{\mu^2}{\sigma^2} - C_2) I = m_t I$. For $0 < t \le t_m$ we have $\mu^2/\sigma^2 \ge 2C_2$ by equation 24 and monotonicity of σ/μ, so $m_t \ge \frac{\mu^2}{\sigma^2} - \frac{1}{2}\frac{\mu^2}{\sigma^2} = \frac{\mu^2}{2\sigma^2} > 0$. Strong convexity yields the unique minimizer and $H \succeq m_t I$. □

Let $\lambda = 1/\tau^2$ be the inverse temperature and write $\langle\cdot\rangle_\lambda$ for expectation under the posterior $q_\lambda(\mathbf{x}_0) \propto e^{-\lambda f_{\mathbf{x}_t}(\mathbf{x}_0)}$; thus $\langle \mathbf{x}_0 \rangle_\tau = \langle \mathbf{x}_0 \rangle_\lambda$. Set $\mathbf{u} = \mathbf{x}_0 - \mathbf{x}_0^*$ and define $\mathbf{b}_\lambda \triangleq \langle \mathbf{u} \rangle_\lambda = \langle \mathbf{x}_0 \rangle_\lambda - \mathbf{x}_0^*$.

The following exact integration-by-parts (Stein) identities hold for any $\lambda > 0$; the rapid decay of q_λ (Lemma B.2) ensures all boundary terms vanish.

Lemma B.3 (Stein identities)**.** *For any $\lambda > 0$,*

$$\langle \nabla_{\mathbf{x}_0} f_{\mathbf{x}_t} \rangle_\lambda = \mathbf{0}, \qquad \langle \mathbf{u} \cdot \nabla_{\mathbf{x}_0} f_{\mathbf{x}_t} \rangle_\lambda = \frac{d}{\lambda}. \tag{28}$$

Proof. Throughout this proof, all gradients are with respect to the integration variable $\mathbf{x}_0$, with $\mathbf{x}_t$ held fixed. For the first identity,

$$\int \nabla_{\mathbf{x}_0} f \, e^{-\lambda f} \, d\mathbf{x}_0 = -\frac{1}{\lambda} \int \nabla_{\mathbf{x}_0} e^{-\lambda f} \, d\mathbf{x}_0 = \mathbf{0},$$

where the last equality follows from the divergence theorem and the rapid decay of $e^{-\lambda f}$ at infinity. For the second identity, integrate by parts componentwise:

$$\int u_i \, \partial_i f \, e^{-\lambda f} \, d\mathbf{x}_0 = \frac{1}{\lambda} \int (\partial_i u_i) \, e^{-\lambda f} \, d\mathbf{x}_0 = \frac{1}{\lambda} \int e^{-\lambda f} \, d\mathbf{x}_0,$$

where $\partial_i u_i = 1$ since $u_i = x_{0,i} - x_{0,i}^*$. Summing over $i = 1, \ldots, d$ and normalizing by $\int e^{-\lambda f} \, d\mathbf{x}_0$ yields $\langle \mathbf{u} \cdot \nabla_{\mathbf{x}_0} f \rangle_\lambda = d/\lambda$. □

(a) The mean shift is controlled by the anharmonicity. Since $\nabla_{x_0} f_{\boldsymbol{x}_t}(\boldsymbol{x}_0^*) = \mathbf{0}$, Taylor's theorem gives $\nabla_{x_0} f_{\boldsymbol{x}_t}(\boldsymbol{x}_0) = H\boldsymbol{u} + \boldsymbol{r}(\boldsymbol{u})$ with $H = \nabla_{x_0}^2 f_{\boldsymbol{x}_t}(\boldsymbol{x}_0^*)$. By $\|\nabla^3 U\| \le C_3$, the remainder satisfies,

$$\|\boldsymbol{r}(\boldsymbol{u})\| \le \tfrac{C_3}{2} \|\boldsymbol{u}\|^2.$$

Taking $\langle\cdot\rangle_\lambda$ of $\nabla f_{\boldsymbol{x}_t} = H\boldsymbol{u} + \boldsymbol{r}(\boldsymbol{u})$ and using equation 28 yields the *exact* identity

$$\boldsymbol{b}_\lambda = -H^{-1} \langle \boldsymbol{r}(\boldsymbol{u}) \rangle_\lambda, \qquad \text{hence} \qquad \|\boldsymbol{b}_\lambda\| \le \frac{1}{m_t} \cdot \frac{C_3}{2} \, \langle \|\boldsymbol{u}\|^2 \rangle_\lambda, \tag{29}$$

where we used $H \succeq m_t I$ so $\|H^{-1}\| \le 1/m_t$.

(b) Second-moment bound. Strong convexity gives $\boldsymbol{u} \cdot \nabla f_{\boldsymbol{x}_t}(\boldsymbol{x}_0) = \boldsymbol{u} \cdot (\nabla f_{\boldsymbol{x}_t}(\boldsymbol{x}_0) - \nabla f_{\boldsymbol{x}_t}(\boldsymbol{x}_0^*)) \ge m_t \|\boldsymbol{u}\|^2$. Averaging and using the second identity in equation 28,

$$m_t \, \langle \|\boldsymbol{u}\|^2 \rangle_\lambda \le \langle \boldsymbol{u} \cdot \nabla f_{\boldsymbol{x}_t} \rangle_\lambda = \frac{d}{\lambda} \quad \Longrightarrow \quad \langle \|\boldsymbol{u}\|^2 \rangle_\lambda \le \frac{d}{\lambda \, m_t} = \frac{d\,\tau^2}{m_t} \;\; (\lambda = 1/\tau^2).$$

(c) Assembling. Combining (a)-(b), for the two temperatures τ (i.e. $\lambda = 1/\tau^2$) and 1,

$$\|\boldsymbol{b}_\tau\| \le \frac{C_3 d\,\tau^2}{2\,m_t^2}, \qquad \|\boldsymbol{b}_1\| \le \frac{C_3 d}{2\,m_t^2}.$$

With the exact reduction $\mathcal{E}_t = \frac{\mu}{\sigma^2}(\boldsymbol{b}_\tau - \boldsymbol{b}_1)$ and the triangle inequality,

$$\|\mathcal{E}_t\| \le \frac{\mu}{\sigma^2}\big(\|\boldsymbol{b}_\tau\| + \|\boldsymbol{b}_1\|\big) \le \frac{\mu}{\sigma^2} \cdot \frac{C_3 d\,(\tau^2+1)}{2\,m_t^2}.$$

Substituting $m_t = \mu^2/\sigma^2 - C_2 = (\mu^2 - C_2\sigma^2)/\sigma^2$ gives the pointwise bound stated in the theorem,

$$\|\mathcal{E}_t\| \le \frac{C_3\, d\,(\tau^2+1)\,\mu(t)}{2\left(\mu(t)^2 - C_2\sigma(t)^2\right)^2}\,\sigma(t)^2. \tag{30}$$

Finally, for $0 < t \le t_m$ we have $C_2\sigma^2 \le \frac{1}{2}\mu^2$ by equation 24, so $\mu^2 - C_2\sigma^2 \ge \frac{1}{2}\mu^2$ and $(\mu^2 - C_2\sigma^2)^2 \ge \frac{1}{4}\mu^4$; together with $\mu(t) \ge \mu(t_m)$, as μ is non-increasing, this turns equation 30 into the uniform bound

$$\|\mathcal{E}_t\| \le \frac{C_3 d\,(\tau^2+1)}{2} \cdot \frac{\mu}{\frac{1}{4}\mu^4}\,\sigma^2 = \frac{2C_3 d\,(\tau^2+1)}{\mu(t)^3}\,\sigma^2 \le \frac{2C_3 d\,(\tau^2+1)}{\mu(t_m)^3}\,\sigma(t)^2 = M\sigma(t)^2,$$

proving equation 26. Note this argument used only C_2, C_3 and strong convexity, no truncation and no expansion parameter–so it is valid on the *entire* interval $(0, t_m]$, not merely asymptotically.

□

B.3 Error Evolution in Reverse SDE

As shown in section B.2, the backward SDE equation 1 shares the same marginal distribution as the forward process equation 5 when omitting the score approximation error. However, in practical sampling, errors arise from not only score error but also discretization error and initial distribution error [Chen et al., 2023]. This section does not model discretization error; it isolates the effect of $\beta(t)$ in continuous time. Intuitively, the error at each step is jointly influenced by the sampling error from the previous step and the network error at the current step. To this end, we introduce three key quantities:

- $p^{(\tau)}$: The exact marginal distribution of the temperature scaled forward process, corresponding to the temperature scaled data distribution at $t = 0$, with $\nabla \log p_t^{(\tau)}$ being the exact score at each step;
- s_θ being the score characterized by the network;
- $\tilde{p}$: The marginal distribution during the actual sampling process, corresponding at $t = T$ to a manually set initial distribution aligning with $p_T^{(\tau)}$ as possible.

For ease of analysis, we reparameterize equation 1 into the following form. Define the reverse time $\tilde{t} = T - t$ ($\tilde{t}$ evolves from 0 to T). Throughout this subsection, all time-dependent coefficients and score functions are evaluated at the corresponding forward time $t = T - \tilde{t}$. For notational simplicity, we write $g = g(T-\tilde{t})$, $\beta = \beta(T-\tilde{t})$, and $s_\theta = s_\theta(x, T-\tilde{t})$, and define $f(x,\tilde{t}) := \frac{\dot{\mu}(T-\tilde{t})}{\mu(T-\tilde{t})}\,x$ where the dot denotes differentiation with respect to the original forward-time variable. Likewise, $p_{\tilde{t}}^{(\tau)}$ and $\tilde{p}_{\tilde{t}}$ denote the corresponding marginal densities reindexed by reverse time. The ideal reverse SDE can then be written as:

$$d\boldsymbol{x}_{\tilde{t}} = -\left[\boldsymbol{f} - \frac{1+\beta}{2}\tau^2 g^2 \nabla_{\boldsymbol{x}} \log p_{\tilde{t}}^{(\tau)}\right] \mathrm{d}\tilde{t} + \tau\sqrt{\beta} g \mathrm{d}\boldsymbol{w}_{\tilde{t}}, \tag{31}$$

and the practical reverse SDE can then be written as:

$$d\boldsymbol{x}_{\tilde{t}} = -\left[\boldsymbol{f} - \frac{1+\beta}{2} g^2 s_\theta\right] \mathrm{d}\tilde{t} + \tau\sqrt{\beta} g \mathrm{d}\boldsymbol{w}_{\tilde{t}}, \tag{32}$$

The true distribution $p_{\tilde{t}}^{(\tau)}$ and the distribution with error $\tilde{p}_{\tilde{t}}$ satisfy their respective Fokker-Planck equations. The former uses the exact score function $\nabla \log p_{\tilde{t}}^{(\tau)}$ in the drift term, while the latter, considering network error, uses s_θ. Therefore:

$$\partial_{\tilde{t}} p_{\tilde{t}}^{(\tau)} = -\nabla \cdot \left(p_{\tilde{t}}^{(\tau)} \left[-\boldsymbol{f} + \frac{1+\beta}{2} \tau^2 g^2 \nabla_{\boldsymbol{x}} \log p_{\tilde{t}}^{(\tau)} \right] \right) + \frac{1}{2} \beta \tau^2 g^2 \Delta p_{\tilde{t}}^{(\tau)}, \tag{33}$$

$$\partial_{\tilde{t}} \tilde{p}_{\tilde{t}} = -\nabla \cdot \left(\tilde{p}_{\tilde{t}} \left[-\boldsymbol{f} + \frac{1+\beta}{2} g^2 s_\theta \right] \right) + \frac{1}{2} \beta \tau^2 g^2 \Delta \tilde{p}_{\tilde{t}}. \tag{34}$$

Using the identity $\Delta p = \nabla \cdot (\nabla p) = \nabla \cdot (p \nabla \log p)$, we can incorporate the Laplace term into the divergence operator, obtaining compact forms:

$$\partial_{\tilde{t}} p_{\tilde{t}}^{(\tau)} = \nabla \cdot \left(p_{\tilde{t}}^{(\tau)} \left[\boldsymbol{f} - \frac{1}{2} \tau^2 g^2 \nabla \log p_{\tilde{t}}^{(\tau)} \right] \right), \tag{35}$$

$$\partial_{\tilde{t}} \tilde{p}_{\tilde{t}} = \nabla \cdot \left(\tilde{p}_{\tilde{t}} \left[\boldsymbol{f} - \frac{1}{2} g^2 s_\theta \right] + \tilde{p}_{\tilde{t}} \left[\frac{\beta}{2} g^2 \left(\tau^2 \nabla \log \tilde{p}_{\tilde{t}} - s_\theta \right) \right] \right). \tag{36}$$

The first term in equation 36 corresponds to the transport of the deterministic probability flow ODE, while the second term introduces a correction proportional to the gradient of the distribution difference $\nabla \log(\tilde{p}_{\tilde{t}} / p_{\tilde{t}}^{(\tau)})$, with its intensity controlled by βg^2.

Now, we examine the evolution of the sampling error characterized by the KL divergence $D_{\mathrm{KL}}(\tilde{p}_{\tilde{t}} \| p_{\tilde{t}}^{(\tau)}) = \int \tilde{p}_{\tilde{t}} \log(\tilde{p}_{\tilde{t}} / p_{\tilde{t}}^{(\tau)}) d\boldsymbol{x}$ over time:

$$\frac{d}{d\tilde{t}} D_{\mathrm{KL}}(\tilde{p}_{\tilde{t}} \| p_{\tilde{t}}^{(\tau)}) = \int (\partial_{\tilde{t}} \tilde{p}_{\tilde{t}}) \log \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} d\boldsymbol{x} + \int \partial_{\tilde{t}} \tilde{p}_{\tilde{t}} d\boldsymbol{x} - \int \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} \partial_{\tilde{t}} p_{\tilde{t}}^{(\tau)} d\boldsymbol{x}. \tag{37}$$

The second term $\int \partial_{\tilde{t}} \tilde{p}_{\tilde{t}} d\boldsymbol{x} = \frac{d}{d\tilde{t}} \int \tilde{p}_{\tilde{t}} d\boldsymbol{x} = 0$ (conservation of probability normalization). Substituting equation 35 and equation 36 and performing integration by parts yields:

$$\begin{aligned}
\frac{d}{d\tilde{t}} D_{\mathrm{KL}}(\tilde{p}_{\tilde{t}} \| p_{\tilde{t}}^{(\tau)}) &= -\int \tilde{p}_{\tilde{t}} \left[\boldsymbol{f} - \frac{1}{2} g^2 s_\theta + \frac{\beta}{2} g^2 \left(\tau^2 \nabla \log \tilde{p}_{\tilde{t}} - s_\theta \right) \right] \cdot \nabla \log \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} d\boldsymbol{x} \\
&\quad + \int \nabla \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} \cdot p_{\tilde{t}}^{(\tau)} \left[\boldsymbol{f} - \frac{1}{2} \tau^2 g^2 \nabla \log p_{\tilde{t}}^{(\tau)} \right] d\boldsymbol{x} \\
&= -\frac{\beta g^2}{2} \int \tilde{p}_{\tilde{t}} \nabla \log \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} \left(\tau^2 \nabla \log \tilde{p}_{\tilde{t}} - s_\theta \right) d\boldsymbol{x} + \frac{g^2}{2} \int \tilde{p}_{\tilde{t}} \nabla \log \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} \left(s_\theta - \tau^2 \nabla \log p_{\tilde{t}}^{(\tau)} \right) d\boldsymbol{x} \\
&= -\frac{\tau^2 \beta g^2}{2} \int \tilde{p}_{\tilde{t}} \| \nabla \log \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} \|^2 d\boldsymbol{x} + (1+\beta) \frac{g^2}{2} \int \tilde{p}_{\tilde{t}} \nabla \log \frac{\tilde{p}_{\tilde{t}}}{p_{\tilde{t}}^{(\tau)}} \left(s_\theta - \tau^2 \nabla \log p_{\tilde{t}}^{(\tau)} \right) d\boldsymbol{x}
\end{aligned} \tag{38}$$

where the integral in the first term is defined as the relative Fisher information $\mathcal{F}(\tilde{p}_{\tilde{t}} \| p_{\tilde{t}}^{(\tau)}) = \int \tilde{p}_{\tilde{t}} \| \nabla \log(\tilde{p}_{\tilde{t}} / p_{\tilde{t}}^{(\tau)}) \|^2 d\boldsymbol{x} \geq 0$.

equation 38 clearly reveals the role of the parameter β. The first term is always non-positive, and strictly negative when a sampling error exists (i.e., $\mathcal{F} > 0$) and $\beta g^2 > 0$, which helps decrease the KL divergence.

The sign of the second term depends on the relationship among the scores of the three distributions: if $(\nabla \log \tilde{p}_{\tilde{t}} - \nabla \log p_{\tilde{t}}^{(\tau)}) \cdot (s_\theta - \tau^2 \nabla \log p_{\tilde{t}}^{(\tau)}) \leq 0$, i.e., the score error vector of the sampling distribution is orthogonal or opposite in direction to the network score error vector, then the second term is also non-positive. In this case, the KL divergence is non-increasing, and strictly decreases when $\tau^2 \beta g^2 \mathcal{F} > 0$.

If the network error is small, the advantage of stochasticity becomes clearer. Define the score approximation discrepancy

$$\delta_{\tilde{t}}(\boldsymbol{x}_t) = s_\theta(\boldsymbol{x}_t, t) - \tau^2 \nabla \log p_{\tilde{t}}^{(\tau)}(\boldsymbol{x}_t), \tag{39}$$

Using the Cauchy-Schwarz inequality, we obtain an upper bound from equation 38:

$$\frac{d}{d\tilde{t}}D_{\mathrm{KL}}(\tilde{p}_{\tilde{t}}\|p_{\tilde{t}}^{(\tau)}) \leq -\frac{\beta\tau^2 g^2}{2}\mathcal{F}(\tilde{p}_{\tilde{t}}\|p_{\tilde{t}}^{(\tau)}) + \frac{(1+\beta)g^2}{2}\sup_{\boldsymbol{x}}\|\delta_{\tilde{t}}(\boldsymbol{x}_t)\|\sqrt{\mathcal{F}(\tilde{p}_{\tilde{t}}\|p_{\tilde{t}}^{(\tau)})}. \tag{40}$$

Therefore, when the network error satisfies

$$\sup_{\boldsymbol{x}}\|\delta_{\tilde{t}}(\boldsymbol{x}_t)\| < \frac{\beta\tau^2}{1+\beta}\sqrt{\mathcal{F}(\tilde{p}_{\tilde{t}}\|p_{\tilde{t}}^{(\tau)})} \tag{41}$$

the KL divergence strictly decreases when $\beta g^2\tau^2\mathcal{F} > 0$. This condition indicates that a larger β imposes a looser upper bound on the network error; when the network error is zero ($\delta_{\tilde{t}} = 0$), a larger β leads to faster decay of the KL divergence.

B.4 Unified Discrete Formulation

Discretization of the Reverse SDE To derive a discrete-time update rule for sampling, we first discretize the reverse SDE in equation 1 using the Euler-Maruyama method. The resulting update step takes the following form:

$$\boldsymbol{x}_{t-\Delta t} = \boldsymbol{x}_t - \left[\frac{\dot{\mu}_t}{\mu_t}\boldsymbol{x}_t - \frac{(1+\beta(t))\tau^2 g_t^2}{2}\nabla_x \log p_t^{(\tau)}(\boldsymbol{x}_t)\right]\Delta t + \sqrt{\beta(t)\tau^2 g_t^2 \Delta t}\boldsymbol{\epsilon}, \tag{42}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$ denotes Gaussian noise.

Assuming locally C^2 schedules, we use first-order Taylor expansions to approximate $\mu_{t-\Delta t}$ and $\sigma_{t-\Delta t}$:

$$\mu_{t-\Delta t} = \mu_t - \dot{\mu}_t\Delta t + O((\Delta t)^2), \quad \sigma_{t-\Delta t} = \sigma_t - \dot{\sigma}_t\Delta t + O((\Delta t)^2) \tag{43}$$

Recalling $g_t^2 = 2(\sigma_t\dot{\sigma}_t - \sigma_t^2\frac{\dot{\mu}_t}{\mu_t})$, and applying the above Taylor expansions, we obtain the approximation:

$$g_t^2\Delta t = 2\sigma_t\dot{\sigma}_t\Delta t - 2\sigma_t^2\frac{\dot{\mu}_t}{\mu_t}\Delta t \approx 2\sigma_t^2\frac{\mu_{t-\Delta t}}{\mu_t} - 2\sigma_t\sigma_{t-\Delta t}. \tag{44}$$

Substituting this result into equation 42 and combining terms yields the final discrete-time update rule:

$$\boldsymbol{x}_{t-\Delta t} = \frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + \frac{(1+\beta(t))\tau^2 g_t^2\Delta t}{2}\nabla_x \log p_t^{(\tau)}(\boldsymbol{x}_t) + \sqrt{\beta(t)\tau^2 g_t^2\Delta t}\boldsymbol{\epsilon}, \tag{45}$$

where $g_t^2\Delta t \approx 2\sigma_t^2\frac{\mu_{t-\Delta t}}{\mu_t} - 2\sigma_t\sigma_{t-\Delta t}$. This formula provides a general discrete-time sampling framework that encompasses a broad family of strategies determined by the specific choice of $\beta(t)$.

Analytic Expectation of the Reverse Probability We first relate the score function to the conditional expectation of the original data:

$$\begin{aligned}
\nabla_{\boldsymbol{x}_t}\log p_t^{(\tau)}(\boldsymbol{x}_t) &= \frac{\int p_0^{(\tau)}(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t}p^{(\tau)}(\boldsymbol{x}_t|\boldsymbol{x}_0)d\boldsymbol{x}_0}{p_t^{(\tau)}(\boldsymbol{x}_t)} \\
&= \frac{\int p_0^{(\tau)}(\boldsymbol{x}_0)p^{(\tau)}(\boldsymbol{x}_t|\boldsymbol{x}_0)(-\frac{\boldsymbol{x}_t-\mu_t\boldsymbol{x}_0}{\tau^2\sigma_t^2})d\boldsymbol{x}_0}{p_t^{(\tau)}(\boldsymbol{x}_t)} \\
&= -\frac{\boldsymbol{x}_t}{\tau^2\sigma_t^2} + \frac{\mu_t}{\tau^2\sigma_t^2}\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t],
\end{aligned} \tag{46}$$

which directly yields the Tweedie formula:

$$\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] = \frac{1}{\mu_t}(\boldsymbol{x}_t + \sigma_t^2\tau^2\nabla_{\boldsymbol{x}_t}\log p_t^{(\tau)}(\boldsymbol{x}_t)) \tag{47}$$

To compute the expectation of the reverse sample $\boldsymbol{x}_{t-\Delta t}$ given $\boldsymbol{x}_t$, we write:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] &= \int \boldsymbol{x}_{t-\Delta t}p^{(\tau)}(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_{t-\Delta t} \\
&= \int \boldsymbol{x}_{t-\Delta t}\int p^{(\tau)}(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t,\boldsymbol{x}_0)p^{(\tau)}(\boldsymbol{x}_0|\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_0\mathrm{d}\boldsymbol{x}_{t-\Delta t} \\
&= \int(\int \boldsymbol{x}_{t-\Delta t}p^{(\tau)}(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t,\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_{t-\Delta t})p^{(\tau)}(\boldsymbol{x}_0|\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_0
\end{aligned} \tag{48}$$

Using equation 11, we know that $\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t, \boldsymbol{x}_0] = \mu_{t-\Delta t}x_0 + \gamma_t \frac{x_t - \mu_t x_0}{\sigma_t}$. Therefore, equation 48 can be further simplified to

$$\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] = \mu_{t-\Delta t}\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] + \gamma_t \frac{x_t - \mu_t \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]}{\sigma_t} \tag{49}$$

Substituting $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ from equation 46, we have

$$\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] = \frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + (\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\tau^2\sigma_t\nabla_x \log p_t^{(\tau)}(\boldsymbol{x}_t), \tag{50}$$

Reverse SDE with Analytic Expectation By aligning the mean term in equation 10 with equation 50, we obtain

$$\begin{aligned}
\frac{(1+\beta(t))\tau^2 g_t^2 \Delta t}{2} &= (\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\tau^2\sigma_t \\
\beta(t)g_t^2\Delta t &= 2(\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\sigma_t - g_t^2\Delta t \\
&\approx 2(\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\sigma_t - (2\sigma_t^2\frac{\mu_{t-\Delta t}}{\mu_t} - 2\sigma_t\sigma_{t-\Delta t}) \\
&= 2\sigma_t(\sigma_{t-\Delta t} - \gamma_t),
\end{aligned} \tag{51}$$

where the approximation in the third line follows from equation 44. Thus the noise coefficient in equation 12 is the first-order schedule expression for $\beta(t)g_t^2\Delta t$, and this calculation provides the continuous-time interpretation of γ_t in terms of $\beta(t)$ in Table 1.

Substituting $g_t^2\Delta t \approx 2\sigma_t^2\frac{\mu_{t-\Delta t}}{\mu_t} - 2\sigma_t\sigma_{t-\Delta t}$, $(1+\beta(t))\sigma_{t-\Delta t} - \beta(t)\sigma_t\frac{\mu_{t-\Delta t}}{\mu_t} = \gamma_t$,

When $\gamma_t = \sigma_{t-\Delta t}$, $\beta(t) = 0$;

When $\gamma_t = \frac{\mu_t\sigma_{t-\Delta t}^2}{\mu_{t-\Delta t}\sigma_t}$, $\beta(t) = \frac{\mu_t\sigma_{t-\Delta t}}{\mu_{t-\Delta t}\sigma_t} \to 1$ as $\Delta t \to 0$;

When $\gamma_t = 0$, $\beta(t) = \frac{2\sigma_t\sigma_{t-\Delta t}}{g_t^2\Delta t} \to \infty$ as $\Delta t \to 0$.

B.5 Sampling with Markov Forward Process Assumption

Proposition B.4 (Parameter γ under the Markov assumption)**.** *Within the linear Gaussian coupling family, assume the forward diffusion process satisfies the Markov property*

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0) = p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t}). \tag{52}$$

Then the associated DDPM-type Gaussian coupling has conditional variance λ_t^2 with

$$\lambda_t = \frac{\sigma_{t-\Delta t}}{\sigma_t}\sqrt{\sigma_t^2 - \frac{\mu_t^2}{\mu_{t-\Delta t}^2}\sigma_{t-\Delta t}^2}, \tag{53}$$

and correspondingly

$$\gamma_t = \sqrt{\sigma_{t-\Delta t}^2 - \lambda_t^2} = \frac{\sigma_{t-\Delta t}^2\,\mu_t}{\sigma_t\,\mu_{t-\Delta t}}. \tag{54}$$

Proof. Following the derivation in equation 11, under a general variance $\tilde{\lambda}_t$, the conditional distribution $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0)$ takes the form

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0) = \mathcal{N}\left(\mu_t\boldsymbol{x}_0 + \tilde{\gamma}_{t-\Delta t}\cdot\frac{\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0}{\sigma_{t-\Delta t}}, \tilde{\lambda}_{t-\Delta t}^2\mathbf{I}\right), \tag{55}$$

where $\tilde{\gamma}_{t-\Delta t} = \sqrt{\sigma_t^2 - \tilde{\lambda}_{t-\Delta t}^2}$. Now consider the identity

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0)\,p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_0) = p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)\,p(\boldsymbol{x}_t \mid \boldsymbol{x}_0), \tag{56}$$

and substitute the Gaussian expressions on both sides. By matching the exponential terms, we obtain the equality

$$\frac{\left\|\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0 - \gamma_t \cdot \frac{\boldsymbol{x}_t - \mu_t \boldsymbol{x}_0}{\sigma_t}\right\|^2}{2\lambda_t^2} + \frac{\|\boldsymbol{x}_t - \mu_t \boldsymbol{x}_0\|^2}{2\sigma_t^2} \tag{57}$$

$$= \frac{\left\|\boldsymbol{x}_t - \mu_t \boldsymbol{x}_0 - \tilde{\gamma}_{t-\Delta t} \cdot \frac{\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0}{\sigma_{t-\Delta t}}\right\|^2}{2\tilde{\lambda}_{t-\Delta t}^2} + \frac{\|\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0\|^2}{2\sigma_{t-\Delta t}^2}. \tag{58}$$

Treating $\boldsymbol{x}_t - \mu_t \boldsymbol{x}_0$ and $\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0$ as independent variables and matching coefficients yields the constraint

$$\frac{\sigma_{t-\Delta t}^2}{\sigma_t^2} = \frac{\lambda_t^2}{\tilde{\lambda}_{t-\Delta t}^2}. \tag{59}$$

Under the Markov assumption, the conditional distribution $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0)$ must be independent of $\boldsymbol{x}_0$. Therefore, from equation 55, the mean term must satisfy

$$\mu_t - \tilde{\gamma}_{t-\Delta t} \cdot \frac{\mu_{t-\Delta t}}{\sigma_{t-\Delta t}} = 0, \tag{60}$$

which implies $\tilde{\gamma}_{t-\Delta t} = \frac{\mu_t \sigma_{t-\Delta t}}{\mu_{t-\Delta t}}$. Substituting into the definition of $\tilde{\lambda}_{t-\Delta t}$, we get

$$\tilde{\lambda}_{t-\Delta t}^2 = \sigma_t^2 - \tilde{\gamma}_{t-\Delta t}^2 = \sigma_t^2 - \left(\frac{\mu_t \sigma_{t-\Delta t}}{\mu_{t-\Delta t}}\right)^2 = \sigma_t^2 - \frac{\mu_t^2}{\mu_{t-\Delta t}^2}\sigma_{t-\Delta t}^2. \tag{61}$$

Finally, using the ratio between λ_t and $\tilde{\lambda}_{t-\Delta t}$, we obtain

$$\lambda_t^2 = \tilde{\lambda}_{t-\Delta t}^2 \cdot \frac{\sigma_{t-\Delta t}^2}{\sigma_t^2} = \left(\sigma_t^2 - \frac{\mu_t^2}{\mu_{t-\Delta t}^2}\sigma_{t-\Delta t}^2\right)\frac{\sigma_{t-\Delta t}^2}{\sigma_t^2}, \tag{62}$$

which is equation 53. The expression for γ_t follows from $\gamma_t = \sqrt{\sigma_{t-\Delta t}^2 - \lambda_t^2}$ and algebraically simplifies to equation 54. □

Proposition B.5 (Continuous BFN sampler and the Markov-type coefficient)**.** *For a continuous Bayesian Flow Network (BFN), let $\boldsymbol{x}_t$ denote the Gaussian input-distribution mean parameter, written in the common marginal form*

$$\boldsymbol{x}_t = \mu_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \sigma_t^2 = \mu_t(1-\mu_t), \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{63}$$

At unit temperature, the standard continuous BFN sampler uses the same Markov-type coefficient, $\gamma_t = \frac{\mu_t \sigma_{t-\Delta t}^2}{\mu_{t-\Delta t}\sigma_t}$,

Proof. The closed-form continuous BFN sampling update [Graves et al., 2023, Xue et al., 2024b], expressed in our reverse-time notation, is

$$\begin{aligned}\boldsymbol{x}_{t-\Delta t} = \frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + \frac{\mu_t - \mu_{t-\Delta t}}{\sqrt{\mu_t(1-\mu_t)}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \\ + \sqrt{\frac{1-\mu_{t-\Delta t}}{1-\mu_t}(\mu_{t-\Delta t} - \mu_t)}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).\end{aligned} \tag{64}$$

We verify both the conditional mean and the conditional variance. Under the reverse-time convention used in equation 64, $\mu_{t-\Delta t} \geq \mu_t$. From equation 63, $\sigma_t = \sqrt{\mu_t(1-\mu_t)}$ and $\sigma_{t-\Delta t} = \sqrt{\mu_{t-\Delta t}(1-\mu_{t-\Delta t})}$. equation 54 therefore gives

$$\gamma_t = \frac{\mu_t\,\mu_{t-\Delta t}(1-\mu_{t-\Delta t})}{\mu_{t-\Delta t}\sqrt{\mu_t(1-\mu_t)}} = \frac{\mu_t(1-\mu_{t-\Delta t})}{\sqrt{\mu_t(1-\mu_t)}}. \tag{65}$$

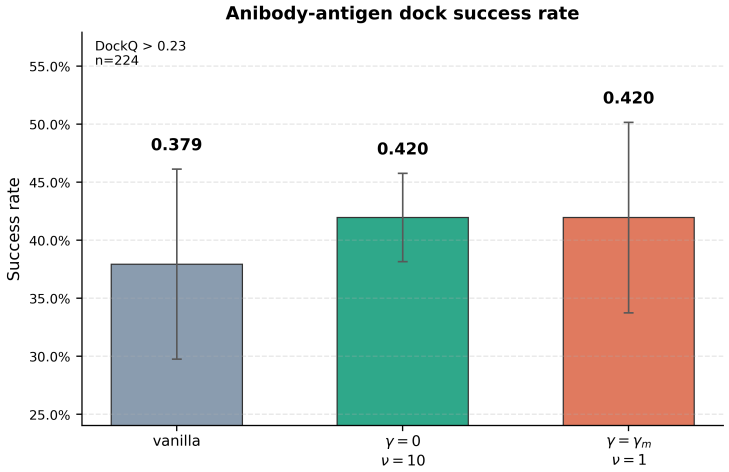


Figure 5: Antibody-antigen docking success rate (DockQ > 0.23) on the evaluation set. Error bars denote standard deviations across samples.

Using the noise-to-score relation $s_\theta(\boldsymbol{x}_t, t) = -\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)/\sigma_t$, the score term in equation 12 becomes

$$\begin{aligned}
&\left(\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t\right)\sigma_t s_\theta \\
&= -\left(\frac{\mu_{t-\Delta t}(1-\mu_t)}{\sqrt{\mu_t(1-\mu_t)}} - \frac{\mu_t(1-\mu_{t-\Delta t})}{\sqrt{\mu_t(1-\mu_t)}}\right)\boldsymbol{\epsilon}_\theta \\
&= \frac{\mu_t - \mu_{t-\Delta t}}{\sqrt{\mu_t(1-\mu_t)}}\boldsymbol{\epsilon}_\theta,
\end{aligned} \tag{66}$$

which is exactly the predicted-noise coefficient in equation 64.

As for the variance term, substituting the BFN schedule into equation 53 yields

$$\begin{aligned}
\lambda_t^2 &= \frac{\mu_{t-\Delta t}(1-\mu_{t-\Delta t})}{\mu_t(1-\mu_t)}\left[\mu_t(1-\mu_t) - \frac{\mu_t^2}{\mu_{t-\Delta t}^2}\mu_{t-\Delta t}(1-\mu_{t-\Delta t})\right] \\
&= \frac{1-\mu_{t-\Delta t}}{1-\mu_t}(\mu_{t-\Delta t} - \mu_t),
\end{aligned} \tag{67}$$

which equals the variance of the final Gaussian term in equation 64. This verifies the conditional-family variance, not the practical first-order variance in equation 12. Hence the continuous BFN sampler is the Markov member of the finite-step conditional family.

□

C Supplementary Experiments

C.1 Antibody-antigen Experiment

We evaluate antibody–antigen docking on the same 224 complexes (138 targets) using DockQ success (DockQ > 0.23). We compare the vanilla churn ODE, StoMax ($\nu = 10$), and Markov sampling ($\nu = 1$). As shown in Figure 5, relative to vanilla, the two variants yield 13 versus 4 and 14 versus 5 discordant failure-to-success versus success-to-failure changes, respectively, indicating modest but consistent observed gains.

C.2 Empirical Verification of Theoretical Claims

Except for the synthetic score-error study below, experiments in this section use QM9 unconditional generation with UniGEM(EDM).

C.2.1 Strong-Convexity Windows under Different Data Distributions and Noise Schedules

The time interval guaranteed by Theorem 2.1 depends jointly on the negative curvature of the data energy and the forward noise schedule. We illustrate this dependence using the one-dimensional symmetric Gaussian mixture

$$p_0(x) = \frac{1}{2}\mathcal{N}(-a, 1) + \frac{1}{2}\mathcal{N}(a, 1), \tag{68}$$

where $a \geq 1$ controls the separation between the two modes. A straightforward differentiation shows $a = 1$ yields a globally convex energy, whereas $a > 1$ produces a non-convex bimodal distribution whose negative curvature increases with the mode separation. To distinguish the exact convexity boundary from the window guaranteed by the theorem, we use two curvature constants. For a general twice-differentiable energy U with bounded Hessian, the tight negative-curvature bound and the Hessian norm bound are defined by

$$C_- := \max\left\{0, -\inf_x \lambda_{\min}\big(\nabla^2 U(x)\big)\right\}, \qquad C_2 := \sup_x \|\nabla^2 U(x)\|_{\mathrm{op}}.$$

Thus $\nabla^2 U(x) \succeq -C_- I$ and $C_- \leq C_2$. The constant C_- determines the exact convexity condition for the auxiliary conditional energy, whereas Theorem 2.1, as stated, uses the two-sided curvature bound C_2.

For the Gaussian mixture considered here,

$$U_a''(x) = 1 - a^2 \operatorname{sech}^2(ax), \qquad \inf_x U_a''(x) = 1 - a^2, \qquad \sup_x U_a''(x) = 1.$$

Consequently,

$$C_-(a) = a^2 - 1, \qquad C_2(a) = \max\{1, a^2 - 1\}.$$

For a fixed noisy state x_t, the auxiliary conditional energy is $f_{x_t}(x_0) = U_a(x_0) + \frac{(x_t - \mu(t)x_0)^2}{2\sigma(t)^2}$. Its curvature satisfies

$$\frac{\partial^2 f_{x_t}(x_0)}{\partial x_0^2} \geq -C_-(a) + \frac{\mu(t)^2}{\sigma(t)^2}. \tag{69}$$

This relation gives rise to two distinct time thresholds. The maximal convexity time is

$$t_\star(a) := \sup\left\{t \in [0,1] : \frac{\sigma(t)}{\mu(t)} \leq \frac{1}{\sqrt{C_-(a)}}\right\}. \tag{70}$$

The threshold $t_\star$ marks the end of the interval on which global convexity is guaranteed. If $\mu(t_\star)^2/\sigma(t_\star)^2 = C_-(a)$, the minimum curvature vanishes at this boundary. Unlike the theorem's threshold t_m, $t_\star$ does not in general guarantee the quantitative strong-convexity margin required by the error bound.

Theorem 2.1 instead uses the more conservative threshold

$$t_m(a) := \sup\left\{t \in [0,1] : \frac{\sigma(t)}{\mu(t)} \leq \frac{1}{\sqrt{2C_2(a)}}\right\}. \tag{71}$$

For every $0 < t \leq t_m$, equation 69 yields $\nabla_{x_0}^2 f_{x_t}(x_0) \geq \frac{\mu(t)^2}{2\sigma(t)^2} > 0$. This quantitative margin is required by the moment bounds in the proof. Hence, $t_m \leq t_\star$: the former is the theorem's guaranteed window, whereas the latter is the maximal convexity boundary. We set $t_\star = 1$ when $C_-(a) = 0$.

For the polynomial schedule associated with Hoogeboom et al. [2022], $\mu(t) = 1 - t^2, \qquad \sigma(t) = \sqrt{1 - (1-t^2)^2}$, equation 71 gives $t_m(a) = \sqrt{1 - \sqrt{\frac{2C_2(a)}{1+2C_2(a)}}}, \qquad a \geq 1.$

For the Karras-EDM schedule of Karras et al. [2022], we use $\mu(t) = 1$ and $\sigma(t) = \left[\sigma_{\min}^{1/\rho} + t\left(\sigma_{\max}^{1/\rho} - \sigma_{\min}^{1/\rho}\right)\right]^\rho$, with $\sigma_{\min} = 0.002$, $\sigma_{\max} = 80$, and $\rho = 7$. Since $\sigma_{\min} > 0$, this practical schedule is a finite-noise approximation to the theorem's $\sigma(0) = 0$ endpoint. Solving $\sigma(t_m) = 1/\sqrt{2C_2(a)}$ gives $t_m(a) = \frac{[2C_2(a)]^{-1/(2\rho)} - \sigma_{\min}^{1/\rho}}{\sigma_{\max}^{1/\rho} - \sigma_{\min}^{1/\rho}}, \qquad a > 1.$

Table 5: The theorem-guaranteed threshold t_m and the maximal convexity boundary $t_\star$ for symmetric Gaussian mixtures. Time is normalized from the data endpoint ($t = 0$) to the maximum-noise endpoint ($t = 1$).

	Schedule from [Hoogeboom et al., 2022]		Schedule from [Karras et al., 2022]	
a	t_m	$t_\star$	t_m	$t_\star$
1.0	0.428	1.000	0.370	1.000
1.2	0.428	0.669	0.370	0.445
1.5	0.394	0.505	0.360	0.393
2.0	0.272	0.366	0.321	0.352

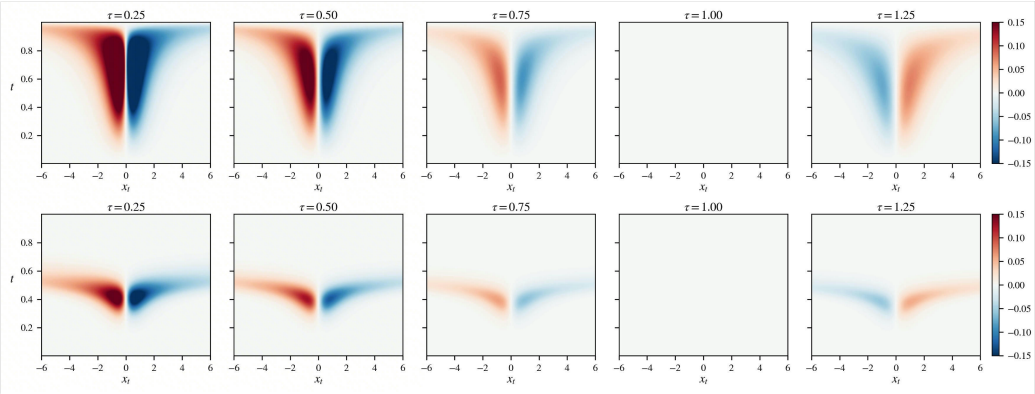


Figure 6: Pointwise score-reuse discrepancy $\nabla_{x_t} \log p_t(x_t) - \tau^2 \nabla_{x_t} \log p_t^{(\tau)}(x_t)$ for the symmetric Gaussian mixture $p_0 = \frac{1}{2}\mathcal{N}(-1.2, 1) + \frac{1}{2}\mathcal{N}(1.2, 1)$. The top and bottom rows use the polynomial and Karras-EDM noise schedules, respectively, as defined in Appendix C.2.1. Columns correspond to the indicated fixed temperatures. The discrepancy decreases toward the low-noise endpoint, consistent with Theorem 2.1, and vanishes identically at $\tau = 1$.

As shown in Table 5, stronger mode separation generally shortens both windows because the quadratic noising term must overcome greater negative curvature. Their values also vary across schedules according to the growth rate of $\sigma(t)/\mu(t)$, so the ordering between schedules need not remain fixed as a changes. The threshold t_m is a sufficient theoretical window rather than an empirically optimized annealing onset: $t_\star > t_m$ shows that convexity can persist beyond the interval on which the proof retains its uniform margin, and the score-reuse approximation may remain accurate even beyond $t_\star$.

C.2.2 Visualization of the Pointwise Score-Reuse Error in Theorem 2.1

We visualize the pointwise score-reuse error estimated in Theorem 2.1 using the one-dimensional mixture $p_0 = \frac{1}{2}\mathcal{N}(-1.2, 1) + \frac{1}{2}\mathcal{N}(1.2, 1)$ together with the polynomial schedule of Hoogeboom et al. [2022] and Karras-EDM schedule of Karras et al. [2022]. Its energy $U = -\log p_0 + \text{const}$ is non-convex, so this example probes the non-log-concave setting covered by the theorem. The exact scores of p_t and $p_t^{(\tau)}$ are computed by numerical integration, so the measured discrepancy reflects solely the theorem's approximation error.

Figure 6 visualizes the signed discrepancy $\nabla \log p_t - \tau^2 \nabla \log p_t^{(\tau)}$ across (x_t, t) for five temperatures. The error peaks at intermediate noise levels and contracts toward the low-noise data endpoint, as implied by the theorem's $O(\sigma(t)^2)$ bound. At $\tau = 1$ the tempered and original distributions coincide exactly, so the discrepancy is identically zero; $\tau < 1$ and $\tau > 1$ yield opposite signs.

Figure 7 further compares the error under three practical power-law annealing schedules $\tau(t) = t^\nu$. Milder schedules ($\nu = 0.3, 0.5$), which keep $\tau \approx 1$ during the high-noise stage and cool only near the data endpoint, produce substantially smaller errors than the aggressive $\nu = 1$ schedule. This provides a direct bridge from the theorem's validity window to the temperature-annealing design used in our experiments.

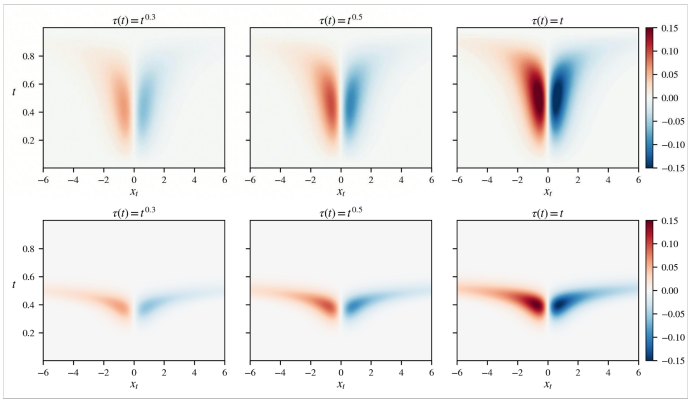


Figure 7: Pointwise score-reuse discrepancy for the same Gaussian mixture under power-law temperature schedules $\tau(t) = t^{\nu}$, with $\nu = 0.3, 0.5, 1$ from left to right. The top and bottom rows use the polynomial and Karras-EDM noise schedules, respectively. In this example, milder cooling produces smaller score-reuse discrepancies over the displayed region.

C.2.3 Temperature Regulation Enhances Distribution Sharpness

The data distribution is inherently sharp while vanilla sampling methods yield smoother distributions; adjusting the temperature can sharpen the main peak of the sampling distribution.

Figure 4 visualizes the bond length distribution of C≡N (triple bond between carbon and nitrogen) in the QM9 dataset. The vanilla DDPM sampling produces a smooth and flat distribution that fails to capture the sharp peak of the real data. As the temperature parameter τ decreases, the sampling distribution progressively sharpens, concentrating more probability mass near the true bond length mode.

To quantify this effect, we compute statistical sharpness metrics including standard deviation (Std) and entropy for bond length distributions. Table 6 presents the results for representative bond types (C-C, C-N, C-O, C=O), comparing annealing exponent $\nu = 0$ (constant schedule, equivalent to vanilla DDPM) with $\nu = 6$ (temperature-annealed schedule).

Table 6: Bond Length Distribution Sharpness Metrics: Effect of Temperature (annealing exponent ν). Lower Std and Entropy indicate sharper distributions (higher peaks).

	Power $n = 0$ (Vanilla)		$\nu = 6$	
Bond	Std (Å)	Entropy	Std (Å)	Entropy
C-C	0.0363	4.768	0.0232	4.085
C-N	0.0512	5.103	0.0321	4.465
C-O	0.0334	4.652	0.0215	4.017
C=O	0.0103	3.097	0.0070	2.447

The results demonstrate consistent trends across all bond types. With lower temperature ($\nu = 6$), the standard deviation decreases by 32-37% for bond length, indicating sharper distributions. The entropy also decreases by 12-21%, confirming the sharpening effect.

Furthermore, we select several representative and common molecular structural features to illustrate the effect of temperature scaling on molecular geometry distributions. As shown in Figure 8, bond-angle distributions (H-C=O, H-C-N, H-N-C, H-O-C, O-C=O) and bond-length distributions (C≡N, C-C, C=C, C=N, C=O) both exhibit strongly peaked profiles in the ground-truth data. The vanilla DDPM ($\nu = 0$) produces noticeably flatter distributions, while increasing ν progressively sharpens the distribution toward the real data mode.

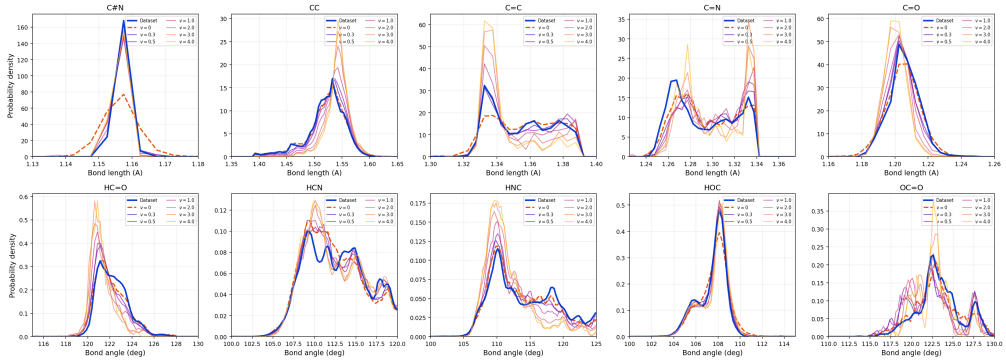


Figure 8: Bond-length distributions (top row) and bond-angle distributions (bottom row) for representative structural features. The vanilla DDPM ($\nu = 0$) produces a smooth distribution, while increasing ν progressively sharpens the distribution toward the real data mode.

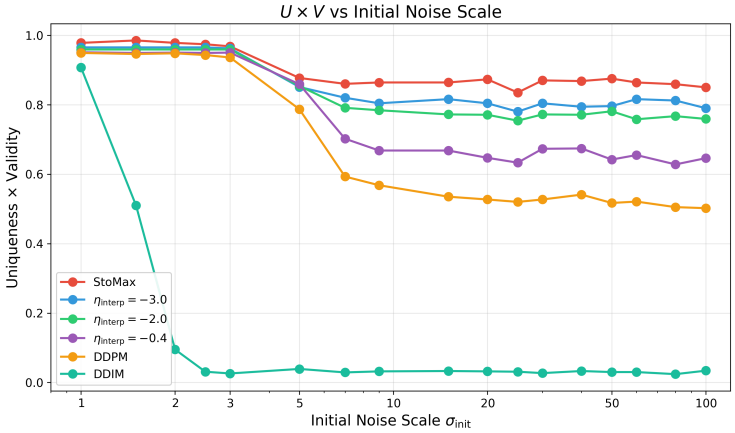


Figure 9: U×V metric across different initial noise scales. 1000 sampling steps. StoMax maintains high performance even at large scales, while DDIM collapses rapidly.

C.2.4 Stochasticity Enhances Robustness to Initial Distribution

We evaluate how different sampling strategies respond to varying initial noise variance. The initial noise scale σ_{init} controls the variance of the starting Gaussian distribution: $\sigma_{\text{init}} = 1.0$ corresponds to standard Gaussian $N(0, 1)$, while larger values represent more dispersed initial distributions. We test 17 scales from 1.0 to 100.0 across six sampling methods.

StoMax achieves the best U×V across all 17 initial noise scales, demonstrating superior robustness. These results show that the stochastic sampler is more tolerant to this initialization perturbation and are consistent with the error-correction perspective of the continuous-time analysis.

C.2.5 Stochasticity Enhances Robustness to Sampling Steps

We evaluate how different sampling strategies perform under reduced sampling budgets. Fewer sampling steps means larger step sizes in the reverse trajectory, which may cause the sampling path to deviate from the optimal trajectory. We test five sampling budgets (1000, 500, 200, 100, 50) across five sampling methods.

StoMax achieves the best U×V at 200 steps (0.972), and notably, StoMax with only 100 steps (U×V=0.959) already surpasses DDPM with the full 1000 steps (U×V=0.949), using **10× fewer denoising steps** with improved quality. This demonstrates finite-step robustness; because Theorem 2.2 excludes discretization error, the result is supporting evidence rather than a direct validation of that theorem.

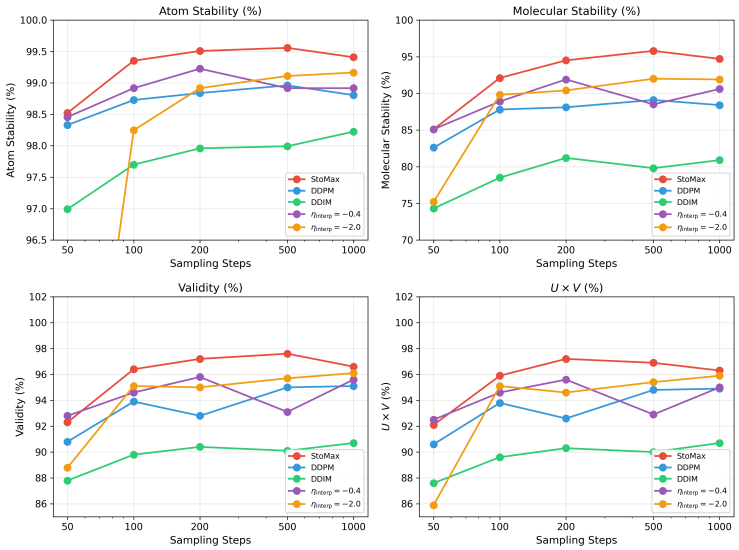


Figure 10: Metrics across different sampling steps. StoMax maintains high performance even with aggressive step reduction, achieving the best results at 100-200 steps.

C.3 Comparative Study of Temperature Schedule Curves

We provide an ablation of the temperature schedule here. The appropriate metric depends on the application: lower temperature is intended to favor physical validity at the potential cost of diversity. While the main text focuses on physical quality, we also evaluate on the U×V quality–diversity trade-off metric, which is more demanding. This analysis compares the empirical behavior of different temperature profiles and supports our choice of the polynomial decay $\tau(t) = t^{\nu}$.

We compare nine constant, decreasing, and increasing temperature schedules (Figure 11) under DDPM and Interp ($\eta_{\text{interp}} = -2$) across eight base temperatures. Full V×U grids are reported in Tables 7 and 8.

Two failure modes are consistent across both samplers. First, schedules that enter the high-noise stage with low temperature i.e., $1 - t^{0.5}$ and sqrt increase, underperform other schedules significantly. Second, schedules that cool too quickly i.e., quadratic decay, cosine annealing, also perform poorly, particularly at low τ_{base}. This pattern is consistent with insufficient early exploration, although the experiment changes entire temperature profiles. Linear increase, which starts from $0.5\tau_{\text{base}}$, performs well despite being an increasing schedule.

Three schedules, i.e. $t^{0.3}$, sqrt decay, and linear decay achieve strong and consistent performance across both samplers at the default $\tau_{\text{base}} = 1$ without requiring fine tuning. We ultimately select the t^{ν} family because it requires only a single exponent while flexibly covering a range of decay rates and curve shapes.

Table 7: V×U (%) for Interp sampling across the nine temperature schedules and 8 base temperatures τ_{base} (1000 samples per cell). Underline marks each schedule's best τ_{base}. Bold font in the Best column marks schedule-level maxima within a descriptive tolerance of 2.8 pp of the global maximum.

Schedule	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0	Best
constant	91.90	95.60	96.60	<u>97.30</u>	95.40	95.90	75.70	69.30	**97.30**
linear decay	81.30	88.20	92.30	94.90	95.40	<u>96.40</u>	88.50	67.70	**96.40**
linear increase	80.00	87.50	92.70	95.50	96.40	<u>96.70</u>	94.90	84.90	**96.70**
sqrt decay	79.70	85.50	89.00	93.20	94.90	<u>96.50</u>	89.30	64.70	**96.50**
sqrt increase	70.50	76.50	79.40	82.20	<u>85.50</u>	85.20	82.30	76.00	85.50
cosine	76.20	80.40	83.80	86.10	89.10	<u>90.80</u>	88.70	56.50	90.80
quadratic	66.50	68.30	71.40	73.60	77.50	79.20	87.20	<u>91.50</u>	91.50
$t^{0.3}$	84.70	89.00	94.20	96.10	<u>96.70</u>	95.50	84.70	66.80	**96.70**
$1 - t^{0.5}$	64.40	65.80	66.20	66.90	68.50	71.50	79.30	<u>85.40</u>	85.40

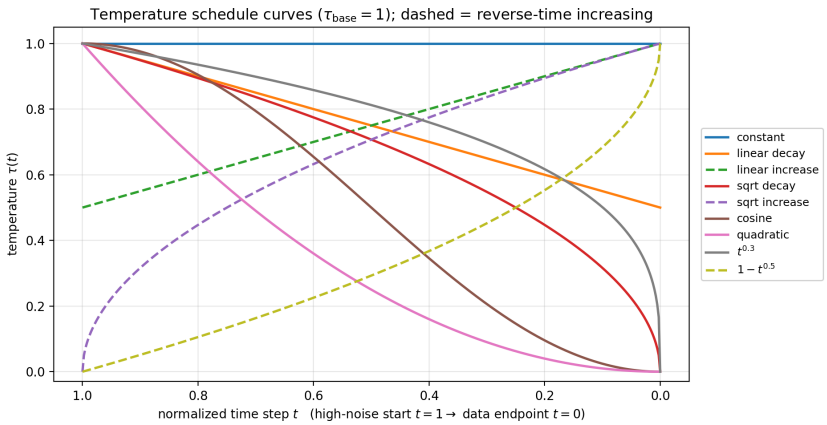


Figure 11: The nine temperature schedule curves compared in this study, drawn at $\tau_{\text{base}} = 1$. The x-axis represents normalized time step t and the y-axis the temperature $\tau(t)$. Here t decreases from 1 at the high-noise start to 0 at the data endpoint, so sampling proceeds left to right; "decay" and "increase" therefore refer to the reverse-time trajectory. Solid lines cool toward the data endpoint, dashed lines heat toward it. Note that heating is not uniformly harmful: linear increase, which only rises from $0.5\tau_{\text{base}}$ to τ_{base}, remains competitive, whereas the schedules that start from $\tau(1) = 0$ and leave the high-noise stage essentially noise-free ($1 - t^{0.5}$ and sqrt increase) are among the weakest in Tables 7 and 8.

Table 8: V×U (%) for DDPM sampling across the nine temperature schedules and 8 base temperatures τ_{base} (2000 samples per cell). Underline marks each schedule's best τ_{base}. Bold font in the Best column marks schedule-level maxima within a descriptive tolerance of 1.7 pp of the global maximum.

Schedule	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0	Best
constant	85.70	92.35	95.75	96.60	95.25	93.85	60.90	29.45	**96.60**
linear decay	77.30	84.75	90.60	93.85	95.50	95.90	75.55	27.75	**95.90**
linear increase	69.25	77.05	85.20	90.60	95.00	95.75	94.45	80.25	**95.75**
sqrt decay	77.25	84.05	90.20	93.80	94.65	96.15	76.20	26.80	**96.15**
sqrt increase	70.45	71.40	75.05	82.05	86.20	90.05	92.80	83.15	92.80
cosine	72.75	81.45	88.50	91.20	94.25	94.90	74.25	22.65	94.90
quadratic	62.05	68.50	73.55	79.15	83.70	88.60	93.80	45.85	93.80
$t^{0.3}$	79.75	87.25	93.20	94.30	96.05	95.50	69.35	29.80	**96.05**
$1 - t^{0.5}$	74.65	74.70	76.35	73.85	73.15	73.15	75.50	85.20	85.20

C.4 Manifold Coverage on QM9

QM9 is an enumerated chemical space containing only approximately 1.3×10^5 molecules with at most nine heavy atoms. Consequently, exact-SMILES uniqueness must decrease with the sample budget even when samples are drawn independently from a distribution with full support. To calibrate this finite-space effect, let q_j be the empirical probability of molecule j in the fixed reference pool used for this comparison, under the same exact-SMILES representation. The expected uniqueness among n i.i.d. draws is

$$\mathbb{E}[\text{Uniq}(n)] = \frac{1}{n} \sum_j \left[1 - (1 - q_j)^n\right]. \tag{72}$$

Table 9 compares this empirical i.i.d. reference with the vanilla sampler and the TREASSURE configuration used in this scaling experiment.

The reference itself falls from 0.951 at 10,000 samples to 0.626 at 100,000. A substantial decline in exact uniqueness is therefore expected from finite-space resampling alone, without any loss of support. The TREASSURE curve follows this decrease rather than separating sharply below the reference. Its exact uniqueness is lower than that of the vanilla sampler at 100,000 samples, as expected when temperature scaling concentrates probability mass, but exact repetition does not determine whether the sampler still covers the molecular manifold.

Table 9: Exact-SMILES uniqueness on QM9 as the sample budget increases. The empirical i.i.d. reference is the analytic expectation in equation 72.

Number of samples	Empirical i.i.d. reference	Vanilla	TREASSURE
10,000	0.951	0.979	0.949
100,000	0.626	0.850	0.694

We evaluate this distinction against the held-out QM9 test split, which is never observed during training. Following the precision-recall framework of Kynkäänniemi et al. [2019], we define molecular neighborhoods using Morgan fingerprints and Tanimoto similarity. We additionally compute the Fréchet ChemNet Distance (FCD) after deduplicating generated molecules, so exact repetitions cannot directly influence the distributional comparison.

Table 10: Manifold quality and held-out coverage on QM9 at 100,000 samples. FCD is computed after exact deduplication.

Method	Precision ↑	Recall ↑	F1 ↑	FCD ↓	Test coverage ↑
Vanilla	0.852	0.948	0.898	0.378	0.290
TREASSURE	0.918	0.917	0.918	0.383	0.357

Precision increases from 0.852 to 0.918, indicating that a larger fraction of generated molecules lies on the real molecular manifold. Recall remains high (0.917 versus 0.948), and the combined F1 score improves from 0.898 to 0.918. Moreover, the deduplicated FCD is essentially unchanged (0.383 versus 0.378), showing that the distinct molecules produced by TREASSURE remain comparably close to the held-out distribution. Held-out test coverage increases by 23% relative, from 0.290 to 0.357, while 41% of generated samples occur in neither the training nor test set. The reduced exact uniqueness therefore reflects controlled concentration and finite-space repetition rather than a comparable loss of manifold coverage, and the out-of-set fraction rules out exact-SMILES training-set memorization and reproduction.

D Supplementary Experimental Settings

Metrics for Small Molecule Generation. We quantify atom stability as the proportion of atoms with correct valency, and molecule stability as the fraction of molecules in which all atoms are stable. Validity is measured using RDKit by verifying whether the generated 3D structures can be successfully converted to SMILES strings. Uniqueness is defined as the fraction of distinct molecules among all valid samples, reflecting exact-sample diversity. Following prior convention, we report V*U, the product of validity and uniqueness, to jointly capture the balance between chemical correctness and diversity. These metrics are intentionally distinguished from fidelity to the empirical training distribution: a tempered sampler changes relative mode weights by design. "Quality" in the small-molecule experiments means chemical stability/validity, while uniqueness and manifold coverage quantify complementary aspects of diversity. Post-hoc force-field relaxation (e.g., UFF/MMFF94) solves a different optimization problem, changes each generated geometry after sampling, and cannot replace conditional adherence or coverage evaluation; our reported method changes the sampling distribution without such relaxation.

Evaluation Set for Protein Structure Prediction We construct a test set from PDB [Berman et al., 2000] entries released after the training cutoff (January 2024), following standard filtering procedures. Specifically, we retain structures released between 2024-01-01 and 2025-07-13, remove polymer chains with fewer than four resolved residues, exclude entries with chain clashes, and filter out structures exceeding 1,000 tokens, 10,000 atoms, or 4.5 Å resolution.

To assess generalization, we restrict evaluation to low-homology chains and interfaces. Predictions are generated for all atoms in each entry, but scoring is performed only on components sufficiently dissimilar from the training data. Low homology is defined as $< 40\%$ sequence identity for polymers and < 0.85 Tanimoto similarity for ligands. Examples are then sampled in a cluster-wise manner to ensure diversity.

Metrics for De Novo Protein Design The evaluation metrics are defined as follows.

(1) Designability: We use ProteinMPNN [Dauparas et al., 2022] to design eight sequences for each structure, which are then refolded using AlphaFold3 with 5 seeds to measure the pLDDT and the RMSD between the refolded structures and the original RFdiffusion3 structures. A designed structure is defined as designable if any of the refolding RMSD is ≤ 2.0 Å.

(2) Rotamer quality: We assess whether the side chains generated by the model adopt common rotamer states. This approach evaluates the side-chain placement quality by benchmarking predicted structures against the Dunbrack rotamer library. For each eligible residue, the method computes backbone dihedral angles (ϕ, ψ) to index the library and measures the predicted χ angles, applying specific mappings for symmetric side chains. It calculates a component log-probability for each rotamer entry using independent circular Gaussians and aggregates these values via log-sum-exp. The resulting probability is normalized into a loss score where lower values indicate better agreement with expected rotamer distributions and higher values correspond to greater deviation.

(3) Diversity: Structural diversity is quantified by computing pairwise TM-scores for all generated structures. These scores are converted into a distance matrix (distance = $1 -$ TM-score) and analyzed using agglomerative clustering across a series of TM-score thresholds from 0.1 to 0.9. At each threshold, diversity is defined as the ratio of distinct clusters to the total number of structures. We integrate these values across all thresholds to compute a "diversity AUC," a scalar metric representing structural variety across all levels of similarity as described in RFdiffusion3[Butcher et al., 2025].

E Related Work

Diffusion Sampling in General Domain. The sampling of diffusion model has been extensively studied in computer vision. Ho et al. [2020] introduced ancestral sampling with a Markovian reverse process, which typically requires hundreds to thousands of denoising steps. Song et al. [2021a] proposed Denoising Diffusion Implicit Models (DDIM), showing that a non-Markovian, deterministic reverse process can generate reasonable samples in tens of steps. DDIM was later shown to correspond to a first-order discretization of the probability flow ODE [Song et al., 2021b], motivating the development of higher-order solvers tailored to diffusion models, such as DPM-Solver [Lu et al., 2022], which exploits the semi-linear structure of the diffusion ODE, and UniPC [Zhao et al., 2023], a unified predictor-corrector framework. Karras et al. [2022] systematically optimized the diffusion design space and proposed a widely adopted second-order Heun solver. Dhariwal and Nichol [2021] introduced classifier guidance by using a separately trained classifier to trade off diversity for fidelity, while Ho and Salimans [2022] proposed classifier-free guidance, which achieves the same effect by interpolating between conditional and unconditional score estimates. Dhariwal and Nichol [2021] also tried low temperature sampling by scaling the Gaussian noise and noise prediction network, but reported producing blurry and smooth images. Despite these advances, the distinct challenge of designing samplers for physically constrained, sharply concentrated distributions, such as molecular geometries, has received little systematic attention.

Temperature Control for Diffusion Models. Temperature control in diffusion models spans three broad families, summarized in Table 11. **(1) Heuristic noise/denoiser scaling** scales the reverse noise or score at inference time [Dhariwal and Nichol, 2021, Ni et al., 2025, Lin et al., 2024, Geffner et al., 2025a]; it is training-free but does not establish that the global endpoint is p_0^{1/τ^2}, and can over-smooth samples for image generation [Dhariwal and Nichol, 2021]. SLDM analogizes the stationary distribution of a Langevin process at a fixed noise level but does not characterize the full temperature-scaled process [Ni et al., 2025]; Proteina modifies the stochastic term and itself notes this is not a principled temperature reduction [Geffner et al., 2025a]. **(2) Gaussian-based score rescaling** derives analytic score-rescaling formulas but only under strong distributional assumptions: Chroma under an isotropic-Gaussian analysis [Ingraham et al., 2023], and Temporal Score Rescaling for mixtures of isotropic Gaussians [Xu et al., 2026]. **(3) Correction-based sampling** achieves principled temperature-scaling targets at the cost of extra computation during sampling: Feynman–Kac correctors derive a weighted SDE for tempered targets q_t^β via the Feynman–Kac formula, and require a particle population with sequential Monte Carlo resampling to correct the sampling distribution along the reverse trajectory [Skreta et al., 2025].

Table 11: Temperature control methods in diffusion models. "Score operation" denotes the inference-time modification to a pretrained score/noise predictor. ✓ = fully satisfied, ✗ = not satisfied, ⊘ = partially satisfied (only under restrictive assumptions).

Method	Score operation	Theoretical guarantee & data assumption	Extra cost
Noise scaling [Dhariwal and Nichol, 2021, Ni et al., 2025, Lin et al., 2024, Geffner et al., 2025a,b]	Scale reverse noise or score	✗ No principled guarantee	✓ None
Chroma [Ingraham et al., 2023]	Rescale pretrained score	⊘ Isotropic-Gaussian analysis only	✓ None
Temporal score rescaling [Xu et al., 2026]	Time-dependent scalar on score	⊘ Well-separated mixture of isotropic Gaussians only	✓ None
FKC[Skreta et al., 2025]	Weighted SDE for q_t^β + SMC particle resampling	✓ Principled annealed target	✗ Simulates K particles for SMC reweighting
TREASSURE	Temper p_0 and σ_t; reuse s_θ	✓ Uniform $O(\sigma^2)$ on $0 < t \leq t_m$ under bounded $\nabla^2 U, \nabla^3 U$; no Gaussian assumption	✓ None

TREASSURE differs from all three families above. **(i)** It separates temperature (τ) and stochasticity (β) into two independently interpretable axes, rather than conflating them as a single noise magnitude. **(ii)** Theorem 2.1 replaces Gaussian or Gaussian-mixture structure with bounded second- and third-order derivatives, allowing anisotropic, multimodal, and non-log-concave densities; its $O(\sigma^2)$ rate motivates late-stage annealing. **(iii)** Theorem 2.2 gives the exact instantaneous KL evolution under score mismatch and identifies when stochasticity strengthens dissipation.

Relation to General Stochastic Frameworks. Stochastic interpolants [Albergo et al., 2024] construct paths between two prescribed endpoint distributions and yield both deterministic and stochastic generative dynamics. Their adjustable diffusion coefficient is closely related to our stochasticity axis β. The objective is nevertheless different: stochastic interpolants provide a training framework for transporting between the prescribed endpoints, whereas TREASSURE is an inference-time modification of a fixed pretrained model whose ideal fixed-temperature target is proportional to p_0^{1/τ^2}. Our contribution is therefore not another diffusion–flow interpolation, but the joint analysis of endpoint tempering, schedule tempering, and inexact-score error within one sampler.

Diffusion-based Molecular Generation and Their Sampling. The rapid development of generative models in computer vision has profoundly influenced 3D molecular generation, with diffusion-based models emerging as the dominant paradigm. Hoogeboom et al. [2022] extended the vanilla DDPM framework [Ho et al., 2020] to 3D atomic coordinates with the E(3) Equivariant Diffusion Model (EDM). Subsequent works have explored alternative generative frameworks adapted to molecular data: EquiFM [Song et al., 2024b] builds on flow matching [Lipman et al., 2023] and leverages deterministic ODE solvers for faster sampling; GeoBFN Song et al. [2024a] unifies continuous-discrete modeling through Bayesian flow networks [Graves et al., 2023]; and SemlaFlow Irwin et al. [2024] employs continuous and discrete flow matching [Campbell et al., 2024] to jointly model atom coordinates and types. Recently, Feng et al. [2025] introduced UniGEM, which adopts a decoupled two-phase design where coordinate diffusion is followed by atom type prediction to mitigate the instability of joint generation. In this work, we apply our TREASSURE sampler on EDM, GeoBFN, and their UniGEM version.

In protein structure prediction, RoseTTAFold [Baek et al., 2021] employs a three-track neural network with iterative recycling mechanism for coordinate refinement, without diffusion modules. RoseTTAFold All-Atom (RFAA) [Krishna et al., 2024] extends RoseTTAFold to generalized all-atom biomolecular modeling and design. Earlier diffusion-based protein design methods such as RFdiffusion [Watson et al., 2023] were built from RoseTTAFold, whereas more recent all-atom

systems such as RF3 [Corley et al., 2025] and RFdiffusion3 [Butcher et al., 2025] are closer to the all-atom modeling branch.

AlphaFold3 [Abramson et al., 2024] uses a diffusion module for atom-coordinate denoising and cites Karras et al. (2022) [Karras et al., 2022] among its methodological references; its released inference pipeline is consistent with a Karras-style sampler, and the supplementary information provides further algorithmic details. Boltz-1 [Wohlwend et al., 2025] and Protenix [Team et al., 2025], as open-source reimplementations of AlphaFold3, adopt diffusion algorithms similar to AlphaFold3. Protenix-mini [Gong et al., 2025, Qiang et al., 2025] replaces the standard stochastic sampling with an efficient few-step ODE sampler to speed up sampling. RoseTTAFold3 [Corley et al., 2025] employs a diffusion-based architecture with explicit chiral and atomic constraints tailored for unified biomolecular modeling, sharing a similar algorithmic framework with AlphaFold3. We replace the AlphaFold3-like sampling in RoseTTAFold3 with our proposed TREASSURE sampler.

For de novo protein design, RFdiffusion uses a DDPM with self-conditioning [Watson et al., 2023]; later variants explore flow matching and motif conditioning [Ahern et al., 2026]. Chroma uses correlated diffusion and derives low-temperature score scaling under an isotropic-Gaussian assumption [Ingraham et al., 2023]. Genie 2 and Proteina instead scale sampling noise as an empirical temperature-like control [Lin et al., 2024, Geffner et al., 2025a], without identifying a global tempered endpoint or its score-reuse error. La-Proteina uses partially latent flow matching [Geffner et al., 2025b]. RFdiffusion3, our baseline, uses an all-atom diffusion model with classifier-free guidance and center-of-mass conditioning [Butcher et al., 2025]; we modify only its sampler.

Boltzmann Generators [Noé et al., 2019] address a complementary problem: learning and reweighting equilibrium conformations of a fixed molecular system so that ensemble probabilities are physically calibrated. Our experiments include de novo molecule/protein generation and structure prediction, where samples need not share one energy function and cross-molecule energy reweighting is not generally defined. TREASSURE instead reweights the learned generative density and makes no claim of recovering a thermodynamically exact equilibrium ensemble.

F Differences to Related Methods

Predictor-Corrector Sampling [Song et al., 2021b] uses an SDE solver as a predictor and additionally employs multiple steps of annealed Langevin dynamics [Song and Ermon, 2019] to correct the current step. This method requires multiple network forward passes per Predictor-Corrector step: one for the Predictor and multiple for the Corrector. By contrast, our method only requires a single network forward pass per iteration and belongs purely to the predictor-only sampling category.

ODE-Churn Sampling AlphaFold3 references Algorithm 2 of [Karras et al., 2022]. The difference is that AlphaFold3 does not use the 2nd-order Heun method but first order method to solve the ODE but retains the churn-type noise perturbation. The AlphaFold3 sampling algorithm written in our notation is:

$$\boldsymbol{x}_{t-\Delta t} = (\boldsymbol{x}_t + \xi_t) + \eta(t - \Delta t - \hat{t})\frac{(\boldsymbol{x}_t + \xi_t) - D_\theta(\boldsymbol{x}_t + \xi_t, \hat{t})}{\hat{t}} \tag{73}$$

where the stochastic term $\xi_t = 1.003\sqrt{\hat{t}^2 - t^2} \cdot \mathcal{N}(\mathbf{0}, I)$, time $\hat{t} = t$, if $t \leq \gamma_{\text{min}}$, else $1.8t$. When the noise magnitude is below the threshold $\gamma_{\text{min}} = 1.0$, i.e., in the final steps of sampling, deterministic sampling is used. $\eta = 1.5$ is the step scale. Here D_θ is the network's prediction of the clean data, which can be equivalently converted to the network's estimate of the score function: $D_\theta = \frac{x_t + \sigma_t^2 s_\theta}{\mu_t}$, where s_θ is the network's estimate of the score function. Under the schedule used by AF3, this relationship can be further simplified: $D_\theta = x_t + t^2 s_\theta$. equation 73 already has the standard Euler direction because $t - \Delta t - \hat{t} < 0$. Substituting $D_\theta(\boldsymbol{x}, \hat{t}) = \boldsymbol{x} + \hat{t}^2 s_\theta(\boldsymbol{x}, \hat{t})$ introduces an additional minus sign, and therefore gives

$$\boldsymbol{x}_{t-\Delta t} = \boldsymbol{x}_t + \xi_t - \eta(t - \Delta t - \hat{t})\hat{t}\, s_\theta(\boldsymbol{x}_t + \xi_t, \hat{t}). \tag{74}$$

In contrast, we integrate deterministic ODE update and stochastic noise perturbation in a unified iterative form derived from the reverse SDE, rather than combining predefined ODE solver and ad-hoc churn noise.

G Limitations

Our method has several limitations. First, TREASSURE can amplify the strengths of a well-trained model but cannot compensate for its fundamental flaws. It sharpens the sampling distribution under the premise that the learned distribution is reasonably accurate, but it cannot correct systematic bias in mode locations, as no additional prior knowledge is introduced beyond the trained score network. We hypothesize that this limitation may be addressed by combining TREASSURE with guidance, prior-informed techniques, or other supplementary methods. Second, there is an inherent trade-off between sample quality and diversity. Lower temperatures suppress minor modes to favor dominant ones, which benefits quality-prioritized tasks but is unsuitable for tasks requiring broad exploration of chemical space and high diversity. Third, the low-temperature sampling theory carries an approximation error that depends on the diffusion schedule, implying that diffusion schedules with faster noise injection during the forward process may be less compatible with TREASSURE's low-temperature sampling. Although our experiments across DDPM, BFN, and Karras-EDM schedules demonstrate broad practicality, this theoretical limitation should be considered when extending the method to new diffusion schedules. Finally, globally bounded derivatives of the statistical energy may fail in singular or strongly repulsive regions of physical molecular landscapes. Our theorem idealizes the locally regular, high-probability region, and a formal localized extension remains future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's core contributions, technical scope, theoretical claims, and experimental findings. All key methodological innovations, theoretical analyses, and empirical performances are clearly outlined and consistent with the full paper, while the generalization boundary and underlying assumptions are appropriately stated.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Justification: We include a standalone limitations section in the appendix, which elaborates the method's assumptions, empirical scope, practical constraints, and unresolved limitations of our approach.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theoretical result is presented with explicit underlying assumptions. Complete and rigorous proofs are included in the appendix, with all statements properly numbered and cross-referenced.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:[Yes]

   Justification: Our approach requires no additional training and adopts open-source pretrained models. We elaborate full experimental settings, evaluation protocols, and pseudocode to ensure reproducibility of all main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We plan to publicly release the code and reproduction scripts.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include dataset descriptions and key hyperparameters. Since our method requires no training, the optimizer is not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes],

   Justification: We provide explicit error bars for all protein structure prediction experiments. For other tasks, we follow established baseline protocols from prior work [Song et al., 2024a, Feng et al., 2025, Geffner et al., 2025a], where error bars are not reported since experimental random seed variation yields negligible performance fluctuations.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our method is training-free, and it does not impose additional computational overhead compared to the baseline sampling method. We have clearly specified the number of experimental steps and sample sizes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed and fully adhere to the NeurIPS Code of Ethics throughout this work. All materials maintain proper anonymity for blind submission, and no content violates ethical guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The work focuses on foundational research in molecular generation and diffusion model optimization, which is primarily theoretical and methodological in nature. As such, the research does not involve direct societal applications or deployments, and thus there are no direct positive or negative societal impacts to discuss.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work focuses on foundational sampling strategies for molecular and protein generation, built upon existing publicly available datasets and pretrained models. No new high-risk datasets or generative models with potential misuse concerns are released in this paper, so no additional safeguards are required.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite all original datasets, baseline models, and related prior works used in this paper. Corresponding references, source information, and usage terms are acknowledged and respected throughout the manuscript.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper introduces no new datasets, models, or other research assets. All experiments are built entirely on existing publicly available resources, so no additional asset documentation is required.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.