Efficient and Unbiased Sampling from Boltzmann Distributions via Variance-Tuned Diffusion Models

Fengzhe Zhang *fz287@cam.ac.uk*
University of Cambridge

Laurence I. Midgley *laurence@angstrom-ai.com*
Ångström AI
University of Cambridge

José Miguel Hernández-Lobato *jmh233@cam.ac.uk*
Ångström AI
University of Cambridge

Reviewed on OpenReview: *https://openreview.net/forum?id=Jq2dcMCS5R*

Abstract

Score-based diffusion models (SBDMs) are powerful amortized samplers for Boltzmann distributions; however, imperfect score estimates bias downstream Monte Carlo estimates. Classical importance sampling (IS) can correct this bias, but computing exact likelihoods requires solving the probability-flow ordinary differential equation (PF–ODE), a procedure that is prohibitively costly and scales poorly with dimensionality. We introduce Variance-Tuned Diffusion Importance Sampling (VT-DIS), a post-training method that adapts the per-step noise covariance of a pretrained SBDM by minimizing the α-divergence ($\alpha = 2$) between its forward diffusion and reverse denoising trajectories. VT-DIS assigns a single trajectory-wise importance weight to the joint forward–reverse process, yielding unbiased expectation estimates at test time with negligible inference-time overhead compared to standard sampling. On the DW-4, LJ-13, and alanine-dipeptide benchmarks, VT-DIS achieves effective sample sizes of approximately 80%, 35%, and 3.5%, respectively, while using only a fraction of the computational budget required by vanilla diffusion + IS or PF-ODE–based IS. Our code is available at `https://github.com/fz920/cov_tuned_diffusion`.

1 Introduction

Sampling from Boltzmann distributions is central to many applications in statistical physics, chemistry, and materials science. Accurate equilibrium samples enable *in silico* prediction of thermodynamic observables, accelerating the discovery of catalysts, drugs, and novel materials and reducing reliance on costly wet-lab experiments. However, the high dimensionality and multimodal nature of realistic energy landscapes render traditional Markov chain Monte Carlo (MCMC) (Metropolis et al., 1953; Hastings, 1970) and molecular dynamics (MD) simulations (Frenkel & Smit, 2023) prohibitively time-consuming for many practical problems.

Normalizing-flow (NF) Boltzmann generators (Noé et al., 2019; Tan et al., 2025; Köhler et al., 2020) amortize sampling by learning an expressive, invertible mapping from a simple base distribution to the target, and exact density evaluation of generated samples permits unbiased importance sampling (IS). However, the requirement of tractable Jacobian determinants constrains architectural flexibility, becoming a bottleneck in high dimensions. Continuous normalizing flows (CNFs) (Chen et al., 2018; Grathwohl et al., 2019) mitigate this issue by parameterizing the transformation as an ordinary differential equation (ODE) and have been combined with E(3)-equivariant architectures for molecular systems (Garcia Satorras et al., 2021; Klein et al., 2023). Yet even CNFs demand expensive Jacobian computations for log-density evaluation.

Score-based diffusion models (SBDMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020) offer an expressive alternative that does not require invertibility. As with any learned sampler, imperfect score estimates bias downstream Monte Carlo estimates. Two IS-based corrections exist for a pretrained score network: (i) solve the probability flow ODE to obtain exact likelihoods—an approach that is computationally intensive, and whose approximations reintroduce bias; or (ii) solve the reverse stochastic differential equation (SDE) and perform IS over the joint forward–reverse trajectory, whose density factorizes into Gaussian terms; however, naively applying this approach typically yields a very low effective sample size (ESS) because the joint proposal distribution cannot fully cover the target distribution.

We adopt the second approach and show that a simple post-training covariance adjustment makes it practical. We introduce **variance-tuned diffusion importance sampling** (VT-DIS), which freezes the score network and optimizes only the per-step noise covariance by minimizing the α-divergence ($\alpha = 2$) between the forward diffusion and reverse sampling processes. VT-DIS yields an unbiased trajectory-wise estimator and achieves high ESS at an orders-of-magnitude lower computational cost than vanilla diffusion + IS or NF-based samplers. Our contributions are:

- **Algorithm.** We propose VT-DIS, a post-training procedure that endows diffusion-based Boltzmann samplers with high-ESS, trajectory-wise importance sampling while incurring negligible inference-time overhead relative to sample generation. Furthermore, we extend VT-DIS to learn both diagonal and full covariance structures, moving beyond isotropic noise.

- **Equivariant extension.** We demonstrate how VT-DIS integrates with E(3)-equivariant diffusion models, enabling the unbiased generation of molecular samples that respect desired symmetry constraints.

- **Empirical validation.** On Gaussian mixtures, toy many-particle systems, and molecular benchmarks (DW-4, LJ-13, and alanine dipeptide), VT-DIS significantly increases ESS relative to naive diffusion + IS with the same number of sampling steps, while incurring only a fraction of the computational cost required for exact likelihood computation via the probability-flow ODE.

2 Related Work

Normalizing-flow (NF) Boltzmann generators learn an invertible mapping from a simple base distribution to the Boltzmann target, enabling exact evaluation of sample densities and unbiased importance sampling via the change-of-variables formula (Noé et al., 2019). Subsequent work improved expressiveness and scalability: Continuous Normalizing Flows (CNFs) based on neural ODEs (Chen et al., 2018; Grathwohl et al., 2018); symmetry-aware flows introduce E(3)-equivariance for molecular systems (Klein et al., 2023). All CNF variants, however, still require computationally expensive Jacobian determinants or ODE integration for exact likelihood evaluation, making unbiased sampling prohibitive for large biomolecules.

Score-based diffusion models learn the noise-perturbed score and generate samples either by reversing the stochastic differential equation (SDE) (Sohl-Dickstein et al., 2015; Ho et al., 2020) or by integrating the probability-flow ODE, which yields exact likelihoods (Song & Ermon, 2019). Although diffusion methods have proven effective for molecular data (Hoogeboom et al., 2022), evaluating proposal log densities via the PF ODE still requires Jacobian-trace computation and becomes expensive in high dimensions.

Beyond the two popular approaches considered above, neural samplers (Vargas et al., 2023a; Richter & Berner, 2023; Vargas et al., 2023b; Zhang & Chen, 2021) can, in principle, generate unbiased samples directly from a target energy function without requiring training data. However, He et al. (2025) show that these methods typically incur substantial computational overhead and are often less efficient than generating training data via MCMC and subsequently fitting a diffusion model. Scalability is also a concern: to the best of our knowledge, no neural sampler currently performs well on molecules in 3D Cartesian coordinates—the focus of this work. We therefore omit neural-sampler baselines from our experiments.

VT-DIS: Bridging the Gap. VT-DIS retains the efficient stochastic reverse trajectory of standard diffusion models while enabling unbiased reweighting. Instead of integrating the ODE, VT-DIS learns per-step Gaussian covariances by minimizing the α-divergence ($\alpha = 2$) between the forward diffusion and reverse sampling processes, thereby maximizing ESS. Concurrent work has tuned covariances under the Kullback–Leibler

divergence to improve sample quality (Ou et al., 2024; Bao et al., 2022); VT-DIS differs by explicitly optimizing the α-divergence objective for importance-sampling efficiency.

3 Background

Before presenting our method, we review three key ingredients: (i) score-based diffusion models (DMs), (ii) their E(3)-equivariant extensions, and (iii) the role of importance sampling (IS) in obtaining unbiased estimators at inference time.

3.1 Score-Based Diffusion Models

Score-based diffusion models (SBDMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020) generate data by first *diffusing* a sample from the data distribution $\pi_{\text{data}}(\boldsymbol{x})$ with progressively increasing Gaussian noise and then *denoising* it in reverse time. The forward process is described by the Itô stochastic differential equation (SDE)

$$d\boldsymbol{x}_t \;=\; \boldsymbol{f}(\boldsymbol{x}_t, t)\,dt \;+\; g(t)\,d\boldsymbol{w}_t, \tag{1}$$

where $t \in [0, T]$ with $T > 0$, $\boldsymbol{f}$ is the drift, g the diffusion coefficient, and $\boldsymbol{w}_t$ a standard Wiener process. Throughout this work we follow the commonly used *variance-exploding* setting of Karras *et al.*(Karras et al., 2022) and set $\boldsymbol{f} \equiv \boldsymbol{0}$ and $g(t) = \sqrt{2t}$. The forward SDE admits two equivalent reverse-time formulations:

$$d\boldsymbol{x}_t \;=\; -2t\,\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\,dt \;+\; \sqrt{2t}\,d\bar{\boldsymbol{w}}_t, \qquad \text{(reverse SDE)} \tag{2}$$

$$d\boldsymbol{x}_t \;=\; -t\,\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\,dt, \qquad \text{(probability-flow ODE)}, \tag{3}$$

where $p_t(\boldsymbol{x}_t)$ is the marginal density at time t and $\bar{\boldsymbol{w}}_t$ is a Wiener process running backwards from T to 0. In practice we approximate the (intractable) score $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)$ with a neural network $s_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ trained by score matching (Hyvärinen & Dayan, 2005; Song & Ermon, 2019).

To draw a sample, we integrate either (2) or (3) backward from $t = T$ to a small cutoff $\varepsilon \approx 0$, and take $\boldsymbol{x}_{\varepsilon}$ as an approximate draw from π_{data}. Solving the reverse SDE with the discrete DDPM sampler (Ho et al., 2020) amounts to the update rule $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_n) = p(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_n, \hat{\boldsymbol{x}}_0(\boldsymbol{x}_n, t_n; \boldsymbol{\theta}))$, where $\hat{\boldsymbol{x}}_0$ is the network's prediction of the clean signal at step n. Given a time grid $\{t_n\}_{n=0}^{N}$, we obtain the full trajectory density $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:N}) = p(\boldsymbol{x}_N) \prod_{n=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_n)$. Both the learned reverse-time transition $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_n)$ and the reference forward kernel $q(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1})$ are Gaussian; their closed-form expressions are provided in Appendix A.3. Alternatively, one may solve the probability-flow ODE (3) with any off-the-shelf ODE solver to obtain samples at $t = \epsilon$.

3.2 E(3)–Equivariant Score-Based Diffusion Models

Standard diffusion models ignore the geometric symmetries that characterise many structured datasets. In molecular systems, for instance, the potential-energy surface is invariant under global *translations*, *rotations*, and *reflections*—the Euclidean group E(3)—as well as permutations of identical atoms. Encoding these symmetries explicitly is critical for producing physically valid and statistically efficient samplers.

Equivariant backbone. Hoogeboom *et al.* (Hoogeboom et al., 2022) introduced E(3)–*equivariant diffusion models* by replacing the standard score network with an Equivariant Graph Neural Network (EGNN) (Satorras et al., 2021). Following the analysis of Xu *et al.* (Xu et al., 2022), a generative process that is equivariant to E(3) transformations must satisfy two conditions:

1. **Rotation and reflection invariance.** For any orthogonal transformation $R \in O(3)$,

$$p(\boldsymbol{x}_N) \;=\; p(R\boldsymbol{x}_N), \quad p(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_n) \;=\; p(R\boldsymbol{x}_{n-1} \mid R\boldsymbol{x}_n),$$

    which implies that the marginal $p(\boldsymbol{x}_n)$ is invariant at every time step.

2. **Translation invariance.** Perfect translation invariance cannot hold for non-degenerate continuous densities; instead, it is imposed by restricting all configurations to the *centre-of-mass (CoM) subspace* $\sum_i \boldsymbol{x}_i = \boldsymbol{0}$, i.e. every sample is translated so that its CoM is at the origin.

Practical implementation. These constraints are enforced by replacing ordinary Gaussians $\mathcal{N}(\cdot)$ with Gaussians projected onto the zero–CoM subspace, denoted $\mathcal{N}_x(\cdot)$ (sampling and density evaluation details are given in Appendix B). Sampling proceeds exactly as in Sec. 3.1: we integrate either the reverse SDE or the probability-flow ODE, but all intermediate states now live in the zero-CoM subspace and all conditional kernels are the projected Gaussians $\mathcal{N}_x$. The resulting pipeline produces molecular configurations that are E(3)-equivariant by construction.

3.3 Diffusion Models with Importance Sampling

Assume that the (unnormalized) target density $\pi_{\text{data}}(\cdot)$ can be evaluated point-wise, whereas drawing i.i.d. samples is infeasible. Directly using raw outputs from a pretrained diffusion model therefore yields biased estimates of expectations $\mathbb{E}_{\pi_{\text{data}}}\big[\phi(\boldsymbol{x})\big]$. IS removes this bias through the weight $\pi_{\text{data}}(\boldsymbol{x})/p_0(\boldsymbol{x})$, where p_0 denotes the diffusion sampler's marginal at $t=0$. Unfortunately, p_0 is typically intractable, and any approximation re-introduces bias. Because we may generate samples either by integrating the reverse SDE or by solving the probability-flow ODE, two corresponding IS strategies arise.

Reverse SDE pathway. When sampling with a DDPM-type discretization of the reverse SDE, define

$$q(\boldsymbol{x}_{0:N}) = \pi_{\text{data}}(\boldsymbol{x}_0) \prod_{n=0}^{N-1} q(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n), \qquad p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:N}) = p(\boldsymbol{x}_N) \prod_{n=0}^{N-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{n+1}), \tag{4}$$

where both kernels are Gaussian and therefore tractable. The trajectory weight $w = \frac{q(\boldsymbol{x}_{0:N})}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:N})}$ yields an unbiased estimator (see Appendix A.3). In practice, however, the mismatch introduced by a finite time discretization causes the effective sample size (ESS) to collapse.

PF ODE pathway. If instead we solve the probability-flow ODE (3), the likelihood of the terminal sample can be written as (Chen et al., 2018; Song et al., 2020)

$$\log p_0\big(\boldsymbol{x}(0)\big) = \log p_T\big(\boldsymbol{x}(T)\big) + \int_0^T \nabla_{\boldsymbol{x}} \cdot \big(-t\, s_{\boldsymbol{\theta}}(\boldsymbol{x}(t), t)\big)\, dt. \tag{5}$$

where $s_{\boldsymbol{\theta}}$ is the score model. Computing the divergence term exactly is prohibitively expensive and is typically approximated with the Hutchinson–Skilling trace estimator (Skilling, 1989; Hutchinson, 1989). Such stochastic estimates, together with time-discretisation error, re-introduce bias and defeat the purpose of IS.

Limitations. Although both pathways are theoretically sound, each faces practical obstacles: the reverse-SDE approach yields vanishing ESS, whereas the ODE approach incurs excessive compute or bias. In the next section we address the first limitation by introducing a covariance-adjusted scheme that dramatically increases ESS in practice.

4 Method

In this section, we introduce the post-training method VT-DIS to improve ESS for diffusion models with importance sampling. We will first give an overview of the whole algorithm and explain each part of the algorithm in more details and the rationale behind them, then we will explain how our method can also be applied to E(3)-Equivariant diffusion models for molecular generations.

4.1 Overview

We build on a pretrained score-based diffusion model, freezing its score network $s_{\boldsymbol{\theta}}$, and introduce a post-training step that adjusts only the per-step noise covariance $\{\Sigma_{\boldsymbol{\phi}}(n)\}_{n=1}^N$, with $\boldsymbol{\phi}$ being learnable parameters. The objective is to make the learned reverse-time process better agree with the true forward (noising) process by minimizing the $\alpha = 2$ divergence, which directly controls the variance of importance weights and hence the ESS. In practice, we: (1) Simulate target trajectories $\{\boldsymbol{x}_{0:N}\}$ from the forward SDE; (2) Compute the

---

Algorithm 1 Post-training Covariance Tuning (VT-DIS)

Require: Score network $s_{\boldsymbol{\theta}}$, initial covariances $\{\Sigma_{\boldsymbol{\phi}}\}$, time grid $\{t_n\}_{n=0}^{N}$, batch size M
Ensure: Tuned covariances $\{\Sigma_{\boldsymbol{\phi}^*}\}$
1: **repeat**
2:     Sample $\{\boldsymbol{x}_{0:N}^{(i)}\}_{i=1}^{M} \sim q(\boldsymbol{x}_{0:N})$
3:     $\ell^{(i)}(\boldsymbol{\phi}) \leftarrow \log q(\boldsymbol{x}_{0:N}^{(i)}) - \log p_{\{\Sigma_{\boldsymbol{\phi}}\}}(\boldsymbol{x}_{0:N}^{(i)})$
4:     $L(\boldsymbol{\phi}) \leftarrow \dfrac{1}{M}\sum_{i=1}^{M}\exp\big(\ell^{(i)}(\boldsymbol{\phi})\big)$
5:     $\{\Sigma_{\boldsymbol{\phi}}\} \leftarrow \{\Sigma_{\boldsymbol{\phi}}\} - \eta\,\nabla L(\boldsymbol{\phi})$
6: **until** convergence
7: $\{\Sigma_{\boldsymbol{\phi}^*}\} \leftarrow \{\Sigma_{\boldsymbol{\phi}}\}$

---

log–density ratio between the forward and reverse joint distributions given current $\{\Sigma_{\boldsymbol{\phi}}(n)\}_{n=1}^{N}$; (3) Update the covariance parameters by minimizing the Monte Carlo estimate of the $\alpha = 2$ divergence; (4) At inference, sample with the tuned covariances and assign each trajectory a single importance weight. We summarize the whole process in Algorithm 1.

4.2 Covariance Parameterization

We explore three complementary ways to parameterize the per-step covariance $\Sigma_{\boldsymbol{\phi}}(n)$.

Isotropic Covariance The standard DDPM proposal $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n)$ uses an isotropic variance $\Sigma_n = \sigma_{\text{ddpm}}^2(n)\boldsymbol{I}$ (where $\sigma_{\text{ddpm}}^2(n)$ is defined in Appendix A.3, (21)). To tune this, for each $n = 1, \cdots, N-1$ we introduce a nonnegative scalar η_n so that

$$\Sigma_{\boldsymbol{\phi}}(n) = \eta_n \sigma_{\text{ddpm}}^2(n)\boldsymbol{I} \tag{6}$$

with $\eta_n \geq 0$ enforced (e.g. via `SoftPlus`) and initialized at $\eta_n = 1$ to recover the standard DDPM baseline. We use $\boldsymbol{\phi}$ to represent all tunable parameters η_n for each n.

Structured covariances To capture per-coordinate anisotropy or correlations, we replace the scalar multiplier with vector or matrix parameters. A diagonal form is

$$\Sigma_{\boldsymbol{\phi}}(n) = \sigma_{\text{ddpm}}^2(n)\,\text{diag}(\eta_{n,1}, \ldots, \eta_{n,d}), \tag{7}$$

with $\eta_{n,i} \geq 0$ and initialized with each $\eta_{n,i} = 1$ so that baseline is recovered. A fully correlated covariance uses a learnable matrix $L_n \in \mathbb{R}^{d \times d}$: $\Sigma_{\boldsymbol{\phi}}(n) = \sigma_{\text{ddpm}}^2(n)\,L_n L_n^{\top}$, with L_n initialized to the identity to recover the baseline. To limit parameter growth which makes optimization hard especially when number of time steps is large, we further propose a practical low–rank plus isotropic decomposition:

$$\Sigma_{\boldsymbol{\phi}}(n) = \sigma_{\text{ddpm}}^2(n)\big(A_n A_n^{\top} + \alpha_n \boldsymbol{I}\big), \tag{8}$$

where $A_n \in \mathbb{R}^{d \times k}$ where $k < d$ is chosen depending on the computing budget and $\alpha_n \geq 0$ are learned, and $A_n A_n^{\top} + \alpha_n \boldsymbol{I}$ is initialized to I by initializing A_n to be zero matrix and $\alpha_n = 1$.

Sample-dependent models In the case that the number of time steps is large and even the low-rank version ((8)) becomes infeasible, we can use a neural network with input t_n to predict the optimal covariance matrix at each step. Additionally, for maximal flexibility, $\Sigma_{\boldsymbol{\phi}}(n)$ can be made a function of the current state $\boldsymbol{x}_n$ (and t_n) via a neural network. That is:

$$\Sigma_{\boldsymbol{\phi}}(n) = \text{NN}_{\boldsymbol{\phi}}(t_n) \quad \text{or} \quad \Sigma_{\boldsymbol{\phi}}(n) = \text{NN}_{\boldsymbol{\phi}}(\boldsymbol{x}_n, t_n)$$

Although this state-dependent form remains valid, our experiments did not show significant ESS gains over the time–only or time-and-sample dependent parameterizations and the additional function evaluation to get covariance increases the computational cost during inference of diffusion models. Therefore, we defer its exploration to future work.

4.3 Extension to E(3)-Equivariant Diffusion

Our VT-DIS covariance tuning readily extends to E(3)-equivariant diffusion models: the only modification is that all trajectories lie in the zero–center–of–mass (CoM) subspace. Isotropic covariances require no change, since it commutes with rotations, reflections, and CoM projection.

For diagonal or fully correlated covariances, the following form of covariance matrix will satisfy E(3)-equivariance and permutation equivariance:

$$\Sigma_\phi(n) \ = \ B_\phi(n) \otimes I_3, \tag{9}$$

where $B_\phi(n) \in \mathbb{R}^{M \times M}$ acts on the M particles, I_3 on the three Cartesian coordinates and $\otimes$ denotes the Kronecker product. The matrices $B_\phi(n)$ are learned per step. This form automatically satisfies E(3)-equivariance (proof in Appendix B). To enforce permutation symmetry, $B_\phi(n)$ must respect the specific particle-identities of the dataset. We illustrate two representative cases below.

Many–particle systems All particles are identical, so a diagonal $B_\phi(n)$ reduces to an isotropic covariance and adds no flexibility. Instead, we parameterize the full covariance

$$B_\phi(n) = (b_n - a_n)\boldsymbol{I} + a_n \boldsymbol{1}\boldsymbol{1}^\top$$

where a_n and b_n are learnable parameters with $b_n > 0$ and $-\frac{b_n}{M-1} < a_n < b_n$ to ensure $B_\phi(n) \succ 0$. Appendix B shows that this choice satisfies both E(3)-equivariance and full permutation invariance, as needed for systems such as n-body systems.

Molecular distributions. Real-world molecules introduce chemical symmetries that are more subtle than the "all–particles–indistinguishable" setting of the above section. For instance, in alanine dipeptide only the three hydrogens bound to the same carbon atom may be permuted freely; the remaining atoms are fixed. To encode such partial symmetries we assign every atom a symmetry label $L_i \in \{1, \ldots, K\}$, $i = 1, \ldots, M$, where atoms sharing the same label are interchangeable while atoms with different labels are not. The covariance $B_\phi(n)$ at sampling step n is then constrained to depend on the pair (i, j) *only* through the labels (L_i, L_j). Proof in Appendix B.3.

1. **Diagonal (per–class variance).** Let $\eta_{n,\ell} \geq 0$ be a learnable variance for class ℓ.

$$B_\phi(n) \ = \ \mathrm{diag}\big(\eta_{n,L_1}, \eta_{n,L_2}, \ldots, \eta_{n,L_M}\big).$$

   Positive definiteness is automatic when every $\eta_{n,\ell} > 0$.

2. **Block-structured (full class–class covariance).** For any learnable matrix $A_n \in \mathbb{R}^{K \times K}$ and a learnable scalar $\alpha_n > 0$. Define

$$\big[B_\phi(n)\big]_{ij} \ = \ (A_n A_n^\top)_{L_i L_j} \ + \ \alpha_n \, \delta_{ij},$$

   where δ_{ij} is the Kronecker delta. The added diagonal term $\alpha_n I$ guarantees $B_\phi(n) \succ 0$.

4.4 Objective Function

Our aim is to align the reverse-process joint proposal $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:N})$ with the forward-diffusion joint $q(\boldsymbol{x}_{0:N})$ (see (4)), in a manner that directly improves importance-sampling performance. Although one could minimize the Kullback–Leibler (KL) divergence $D_{\mathrm{KL}}(q\|p_\phi)$ (and indeed closed-form optimal covariances under this criterion have been derived (Ou et al., 2024)), such minimization does not control tail mismatch. Poor alignment in the tails leads to high variance of importance weights and thus low ESS when sample size is large. Instead, we minimize the α-divergence with $\alpha = 2^1$, which is proportional to the variance of importance weights:

$$\arg\min_\phi D_{\alpha=2}\big(q\|p_\phi\big) = \arg\min_\phi \left(\mathcal{L}(\phi) := \int \frac{q(\boldsymbol{x}_{0:N})}{p_\phi(\boldsymbol{x}_{0:N})} q(\boldsymbol{x}_{0:N}) \, \mathrm{d}\boldsymbol{x}_{0:N} \approx \frac{1}{M} \sum_{i=1}^{M} \frac{q(\boldsymbol{x}_{0:N}^{(i)})}{p_\phi(\boldsymbol{x}_{0:N}^{(i)})}\right), \quad \boldsymbol{x}_{0:N}^{(i)} \sim q \quad (10)$$

---

[1]We use the α-divergence $D_\alpha\big(q\|p_\phi\big) = \frac{1}{\alpha(\alpha-1)}\big(\int q(x)^\alpha \, p_\phi(x)^{1-\alpha} \, dx - 1\big)$ $(\alpha \neq 0, 1)$. In our experiments we minimize the case $\alpha = 2$; the optimized loss $\mathcal{L}(\phi)$ is the integral term above.

In practice, we draw M trajectories $\{\boldsymbol{x}_{0:N}^{(i)}\}$ from the forward process[2] q to form the unbiased Monte Carlo estimate and minimize $\log \mathcal{L}(\phi)$ for numerical stability (see Appendix C.1). Optimizing this loss by gradient descent yields covariance parameters that minimize importance weight variance and thereby maximize ESS.

4.5 Choice of Sampling Schedule

Karras et al. (2022) propose a family of continuous-time schedules controlled by a parameter ρ:

$$t_n \;=\; \left(t_N^{1/\rho} + \tfrac{n}{N-1}\left(t_0^{1/\rho} - t_N^{1/\rho}\right)\right)^\rho, \qquad t_0 = \epsilon,\; t_N = T.$$

In the limit $\rho \to \infty$, this reduces to a fixed geometric grid $t_n = t_0 \left(t_N/t_0\right)^{n/(N-1)}$, which we found consistently yields lower $\alpha = 2$ divergence and higher ESS across all benchmarks. We also experimented with jointly tuning the schedule and per-step covariances during VT-DIS. While joint tuning improves performance on simple Gaussian mixtures, it proved unstable on molecular datasets. We defer a robust treatment of schedule learning to future work.

4.6 Complexity & Inference-Time Cost

Relative to a standard DDPM equipped with importance sampling, VT-DIS adds only a *single, offline* optimization step: tuning the per-step noise covariances $\{\Sigma_\phi(n)\}$. After this one-time cost, generate samples and importance weights only adds negligible overhead to standard sampling from diffusion models—no extra function evaluations. Unlike flow-based Boltzmann generators or diffusion generate samples by solving PF ODE, VT-DIS never has to integrate an ODE to evaluate marginal proposal log-densities, a procedure that becomes prohibitively expensive in high-dimensional settings.

5 Experiments

Experimental setup. We assume access to both the energy function and the training data used to fit the diffusion model (same assumption as in, e.g., Klein et al. (2023); Tan et al. (2025)). We first train a diffusion model with an EGNN backbone (Satorras et al., 2021) and EDM preconditioning (Karras et al., 2022) using a standard training procedure. We then apply VT-DIS to the same training set to optimize the proposal covariances. For evaluation, we compare VT-DIS with a DDPM+IS baseline under a fixed compute budget measured by the number of function evaluations (NFEs; i.e., sampling steps). We also compare against importance sampling with exact likelihoods computed via the probability-flow ODE (PF-ODE). Performance is reported as both forward and reverse effective sample size (ESS). We additionally report the Evidence Lower Bound (ELBO) and Evidence Upper Bound (EUBO) for the DDPM baseline and VT-DIS, which serve as complementary likelihood-based metrics to assess the quality of the generated samples.

We consider three representative classes of Boltzmann systems: (i) **Synthetic**: a high-dimensional two-mode Gaussian mixture (GMM-2); (ii) **Many-particle**: double-well (DW-4) and Lennard-Jones-13 (LJ-13) clusters (Köhler et al., 2020); and (iii) **Small-molecule**: alanine dipeptide (ALDP). For all targets, the (unnormalized) density π_{data} is available, enabling exact importance weights. We use the atomic coordinate representation for each sample and train our models on these coordinates. Further implementation details are provided in Appendix C.

5.1 GMM-2

We begin with a two-component isotropic Gaussian mixture in $\mathbb{R}^d$:

$$\pi_{\text{data}}(\boldsymbol{x}) = w_1\,\mathcal{N}\!\left(\boldsymbol{x};\,\boldsymbol{\mu}_1, \sigma_1^2\boldsymbol{I}\right) + w_2\,\mathcal{N}\!\left(\boldsymbol{x};\,\boldsymbol{\mu}_2, \sigma_2^2\boldsymbol{I}\right),$$

where $\boldsymbol{\mu}_1 = (1,\dots,1)^\top \in \mathbb{R}^d$, $\boldsymbol{\mu}_2 = (-2,\dots,-2)^\top \in \mathbb{R}^d$, $\sigma_1^2 = \sigma_2^2 = 0.15$, and $w_1 = \tfrac{2}{3}$, $w_2 = \tfrac{1}{3}$. We evaluate two ambient dimensions, $d \in \{50, 100\}$. Figure 1 shows forward and reverse ESS as a function of the number of

---

[2]Empirically, we found that attempting to estimate D_2 with samples from the learned reverse process p_θ often induces mode collapse and degrades forward ESS.
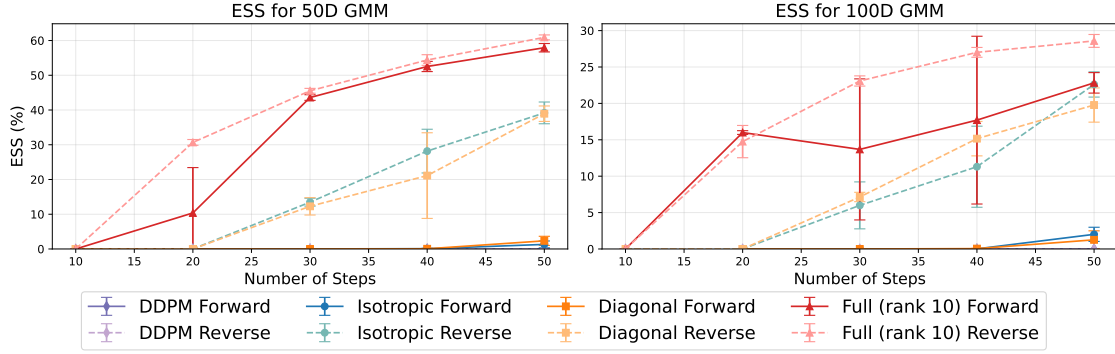
Figure 1: **GMM-2**: ESS versus NFEs for dimensions $d = 50$ and $d = 100$, comparing VT-DIS with the DDPM + IS baseline. ESS estimates are computed using 10^5 samples.

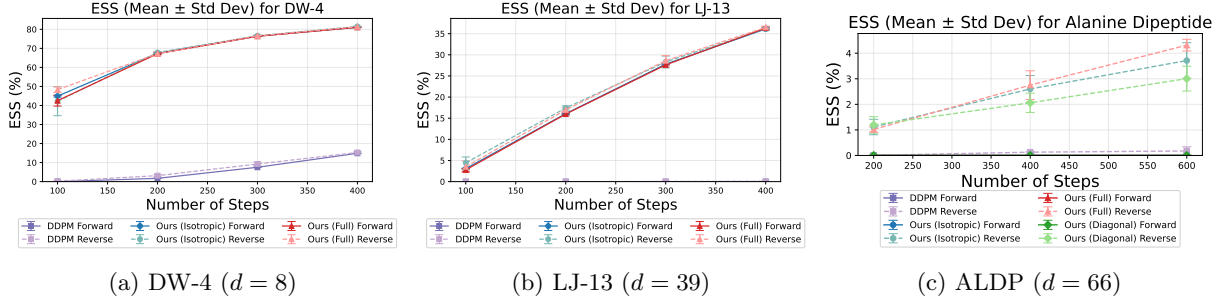(a) DW-4 ($d = 8$)   (b) LJ-13 ($d = 39$)   (c) ALDP ($d = 66$)

Figure 2: **ESS versus NFEs.** VT-DIS consistently outperforms the standard DDPM + IS baseline across all tasks under the same computational budget. ESS values are estimated using 10^5 samples.

function evaluations (NFE) for VT-DIS with three noise–covariance parameterizations—*isotropic*, *diagonal*, and *full rank*—together with the DDPM + IS baseline.

The plots reveal three key trends: (i) VT-DIS consistently outperforms the baseline in forward ESS, reverse ESS, or both, for every covariance parameterization. (ii) The full-rank covariance yields the highest forward and reverse ESS for both values of d. For $d = 100$, however, the forward-ESS error bars are wide when $d = 100$, indicating residual mismatch between the forward and reverse processes at very high dimension. (iii) With isotropic and diagonal covariances, reverse ESS improves as NFE grows, whereas forward ESS stays low, underscoring the importance of modeling cross-dimensional correlations in this setting.

5.2 Many–particle systems: DW-4 and LJ-13

We next evaluate VT-DIS on two prototypical many-body systems with pair-wise interactions: *DW-4* (double-well potential, four particles in $\mathbb{R}^2$, $d = 8$) and *LJ-13* (Lennard–Jones cluster, thirteen particles in $\mathbb{R}^3$, $d = 39$)(Köhler et al., 2020).

ESS versus NFEs. Figure 2 reports forward and reverse ESS as a function of the number of NFEs. On the low-dimensional DW-4 system, VT-DIS raises both ESS measures from $\leq 1\%$ to $\approx 40\%$ using only 100 NFEs. The improvement is even more pronounced on the higher-dimensional LJ-13 cluster: the standard DDPM + IS baseline yields virtually 0% ESS up to 400 NFEs, whereas VT-DIS achieves $\approx 35\%$ ESS on both forward and reverse measures with the same budget. Due to the permutation-equivariant constraint of the many-particle system, the full covariance matrix has only one additional degree of freedom compared to an isotropic covariance. In this setting, we observe no significant ESS gain from using the full covariance, although both forward and reverse ESS exhibit lower standard deviations than with isotropic noise. Across both tasks, ESS increases monotonically with the number of NFEs, facilitating a clear accuracy–cost trade-off.

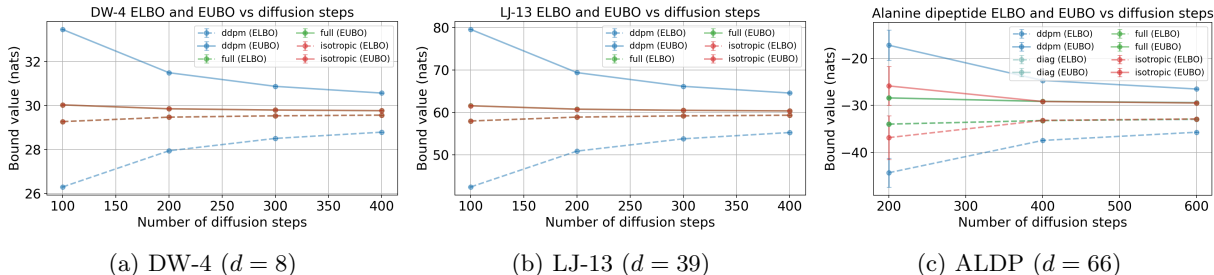(a) DW-4 ($d = 8$)   (b) LJ-13 ($d = 39$)   (c) ALDP ($d = 66$)

Figure 3: **ELBO/EUBO vs. diffusion steps.** ELBO (higher is better) and EUBO (lower is better) for the DDPM baseline and VT-DIS (isotropic, diagonal and full covariances) as a function of the number of sampling steps. Error bars denote $\pm$ one standard deviation across 3 repeated MC evaluations.

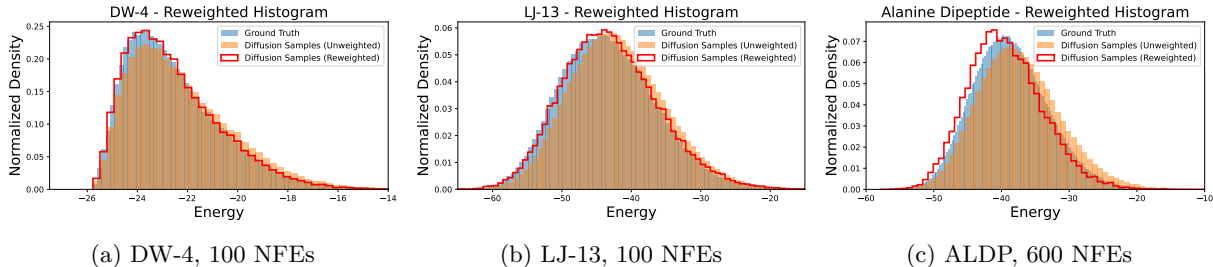(a) DW-4, 100 NFEs   (b) LJ-13, 100 NFEs   (c) ALDP, 600 NFEs

Figure 4: **Energy distributions.** Unweighted samples generated by VT-DIS with isotropic covariance (orange) are biased relative to ground-truth samples (blue), but importance-weighted samples (red) effectively correct the discrepancy. Histograms are plotted using 5×10^5 samples.

ELBO and EUBO. We also report the standard likelihood bounds (ELBO $\leq \log p_\theta(\boldsymbol{x}) \leq$ EUBO), where a smaller gap indicates better agreement between the forward noising path and the learned reverse process. Figure 3 shows that across DW-4 and LJ-13, VT-DIS consistently raises ELBO and lowers EUBO relative to the DDPM baseline, yielding a markedly tighter "sandwich" at every step budget. Overall, these trends mirror the ESS gains, indicating better forward–reverse matching under fixed compute.

Bias correction. To illustrate bias correction we compare three energy histograms for each task: (i) *ground-truth* samples from the test dataset, (ii) *unweighted* VT-DIS generated samples, and (iii) *reweighted* samples. Figure 4 shows that the raw proposals are biased, but importance reweighting realigns the distributions with the target, corroborating the ESS improvements.

5.3 Small Molecule: Alanine Dipeptide (ALDP)

We evaluate VT-DIS on the 66-dimensional Boltzmann distribution of alanine dipeptide (ALDP) in implicit solvent at temperature $T = 800$ K. The unnormalized Boltzmann density is defined by a classical force-field potential, and the availability of its exact log density enables unbiased importance-sampling evaluation. Training and implementation details are provided in Appendix C.

Results. Figure 2c shows the forward and reverse ESS over a range of 200–600 NFEs. VT-DIS consistently improves the reverse ESS, reaching approximately 2.5% at 400 NFEs, compared to under 0.5% for the baseline. Among the covariance parameterizations, the full covariance version achieves slightly higher reverse ESS compared to diagonal and isotropic parameterizations. Figure 3c further illustrates that VT-DIS with all three parameterizations yields a smaller gap between ELBO and EUBO relative to the DDPM baseline. Notably, the full covariance parameterization exhibits an even smaller ELBO–EUBO gap in the low diffusion-step regime, which aligns with the observed ESS performance.

Figure 4c displays the energy histogram, and Figure 5 presents the Ramachandran plot. Both demonstrate that importance sampling effectively corrects bias in the generated samples. Notably, the molecular dynamics
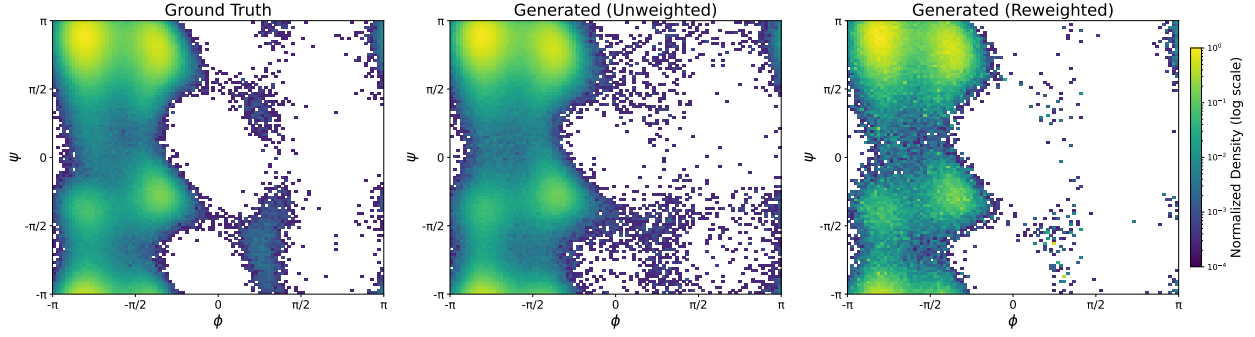
Figure 5: Ramachandran plots from left to right: ground-truth samples generated by MD, VT-DIS (isotropic) samples before and after reweighting. All plots are generated using 10^6 samples.

(MD) "ground-truth" samples themselves exhibit a slight bias, as the reweighted energy histogram shifts toward lower energies than the MD samples.

For both VT-DIS and the baseline, the forward ESS is nearly zero due to a small number of extreme density-ratio outliers. This indicates that both methods' joint proposal distributions assign extremely low probability to regions containing true samples from the test dataset. We hypothesize that the score model is missing certain modes, supported by three empirical observations: (i) in the unreweighted samples (middle plot of Figure 5), the mode near $\phi = \pi/2$ is underrepresented; (ii) running the DDPM baseline with very large numbers of time steps (1000 and 2000) still resulted in forward ESS ≈ 0, indicating a support-coverage mismatch rather than discretization issues; (iii) subsequent experiments solving the PF ODE for importance sampling (Table 1) similarly yielded forward ESS ≈ 0 on Alanine Dipeptide.

Since the score model defines the mean of the per-step proposal distributions and VT-DIS only adjusts the covariance, its capacity to improve *forward ESS* is limited if the score model is inadequately trained. Therefore, VT-DIS yields substantial improvement of forward ESS only when the score model sufficiently captures all modes. Two promising future directions include: (i) enhancing the training of the score model to better capture all relevant modes—a direction complementary to our current work, as VT-DIS can enhance any trained diffusion model; and (ii) extending VT-DIS to adjust both the mean and covariance of the per-step proposal distributions, particularly useful when the score model is missing certain modes.

5.4 Computational Cost versus Probability–Flow ODE

While VT-DIS and PF-ODE + IS both yield unbiased estimators, they differ substantially in computational efficiency. Here, we compare the ESS and computational cost of each method.

Experimental setup. To isolate inference cost, we reuse the *same* pretrained score network and employ an identical number of integration steps for both methods. Computational effort is measured in *giga floating-point operations (GFLOPs) per sample*, and accuracy is quantified by forward and reverse ESS. Results for DW-4, LJ-13, and ALDP are summarized in Table 1. We also include an extended discussion and results on solving the PF-ODE using the trace estimator in Appendix D.1.

Results. PF-ODE + IS attains higher ESS but incurs a steep computational penalty: its GFLOPs per sample are one to two orders of magnitude greater than VT-DIS's, and this gap widens with dimensionality. In practice, PF-ODE requires a *minimum* number of ODE steps to control proposal-density bias; each additional step further increases cost. VT-DIS, in contrast, trades some ESS for dramatically lower compute. Its estimator is asymptotically unbiased as the sample number increases, regardless of the time-step budget, and each additional reverse-SDE step directly boosts ESS.

Table 1: Performance on DW-4, LJ-13 and ALDP with 10^4 samples on a single A100. The number in parentheses after each dataset denotes the sampling steps. $\uparrow$ = higher is better, $\downarrow$ = lower is better.

	DW-4 (100)		**LJ-13** (200)		**ALDP** (200)	
	Rev. SDE	PF ODE	Rev. SDE	PF ODE	Rev. SDE	PF ODE
Forward ESS (%) $\uparrow$	44.32	94.29	15.87	79.21	≈ 0	≈ 0
Reverse ESS (%) $\uparrow$	46.76	83.60	17.69	77.42	1.05	35.39
GFlops / sample $\downarrow$	0.81	7.70	26.2	2173	33.7	45967

6   Conclusion and Limitations

Contributions. We have introduced VT-DIS, a post-training procedure that can be applied to any pre-trained diffusion model, and demonstrated its effectiveness in molecular generation. By computing trajectory-wise importance weights, VT-DIS achieves high effective sample size (ESS) across multiple benchmarks while using only a fraction of the computational budget required to solve the PF ODE for exact likelihood estimation. Furthermore, VT-DIS provides asymptotically unbiased expectation estimates as the number of samples grows—independent of the number of sampling steps. In contrast, PF-ODE solvers guarantee asymptotically unbiased sampling only when both the number of samples and the number of discretization steps increase, which limits their practical utility. These findings highlight VT-DIS's potential for diffusion-based Boltzmann generators.

Limitations and Future Work. Despite its advantages, VT-DIS has several limitations. First, the post-training procedure must be repeated for each sampling schedule, which can incur significant computational and memory costs. Figure 6 in Appendix D.2 shows a clear pattern in optimal hyperparameters across different step counts; future work could leverage this pattern to develop a unified post-training algorithm that supports arbitrary schedules in one stage. Second, optimizing the sampling schedule itself to minimize the α-divergence remains an open question; existing methods from (Sabour et al., 2024; Williams et al., 2024) may offer a promising starting point. Third, VT-DIS requires access to ground-truth samples to tune the per-step proposal covariances. When such samples are unavailable, a dataset must first be generated—e.g., via MCMC or molecular dynamics (MD). A useful direction for future work is to adapt VT-DIS to operate with only a small set of approximate reference samples, which are easier to obtain.

Acknowledgments

References

Tara Akhound-Sadegh, Jarrid Rector-Brooks, Avishek Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, et al. Iterated denoising energy matching for sampling from boltzmann densities. *arXiv preprint arXiv:2402.06121*, 2024.

Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. *arXiv preprint arXiv:2206.07309*, 2022.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2023.

Victor Garcia Satorras, Emiel Hoogeboom, Fabian Fuchs, Ingmar Posner, and Max Welling. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems*, 34:4181–4192, 2021.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, volume 3, 2019.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Jiajun He, Yuanqi Du, Francisco Vargas, Dinghuai Zhang, Shreyas Padhy, RuiKang OuYang, Carla Gomes, and José Miguel Hernández-Lobato. No trick, no treat: Pursuits and challenges towards simulation-free training of neural samplers. *arXiv preprint arXiv:2502.06685*, 2025.

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

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

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

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.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36:59886–59910, 2023.

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

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Laurence Midgley, Vincent Stimper, Javier Antorán, Emile Mathieu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Se (3) equivariant augmented coupling flows. *Advances in Neural Information Processing Systems*, 36:79200–79225, 2023.

Laurence Illing Midgley, Vincent Stimper, Gregor NC Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. *arXiv preprint arXiv:2208.01893*, 2022.

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.

Zijing Ou, Mingtian Zhang, Andi Zhang, Tim Z Xiao, Yingzhen Li, and David Barber. Improving probabilistic diffusion models with optimal covariance matching. *arXiv preprint arXiv:2406.10808*, 2024.

Lorenz Richter and Julius Berner. Improved sampling via learned diffusions. *arXiv preprint arXiv:2307.01198*, 2023.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. *arXiv preprint arXiv:2404.14507*, 2024.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

John Skilling. The eigenvalues of mega-dimensional matrices. *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*, pp. 455–466, 1989.

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

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Charlie B Tan, Avishek Joey Bose, Chen Lin, Leon Klein, Michael M Bronstein, and Alexander Tong. Scalable equilibrium sampling with sequential boltzmann generators. *arXiv preprint arXiv:2502.18462*, 2025.

Francisco Vargas, Will Grathwohl, and Arnaud Doucet. Denoising diffusion samplers. *arXiv preprint arXiv:2302.13834*, 2023a.

Francisco Vargas, Shreyas Padhy, Denis Blessing, and Nikolas Nüsken. Transport meets variational inference: Controlled monte carlo diffusions. *arXiv preprint arXiv:2307.01050*, 2023b.

Christopher Williams, Andrew Campbell, Arnaud Doucet, and Saifuddin Syed. Score-optimal diffusion schedules. *arXiv preprint arXiv:2412.07877*, 2024.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

Yilun Xu, Shangyuan Tong, and Tommi Jaakkola. Stable target field for reduced variance score estimation in diffusion models. *arXiv preprint arXiv:2302.00670*, 2023.

Qinsheng Zhang and Yongxin Chen. Path integral sampler: a stochastic control approach for sampling. *arXiv preprint arXiv:2111.15141*, 2021.

A Score-based Diffusion Models with Importance Sampling

In this appendix, we present the mathematical derivations for combining score-based diffusion models with importance sampling. We first review the fundamentals of importance sampling and then show how it can be integrated into score-based diffusion to correct bias in generated samples.

A.1 Review of Importance Sampling

Importance sampling (IS) provides an estimator for expectations of the form

$$\mathbb{E}_{\boldsymbol{x} \sim \pi_{\text{data}}}[\phi(\boldsymbol{x})] = \int \phi(\boldsymbol{x})\, \pi_{\text{data}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},$$

where π_{data} is the (possibly unnormalized) target density and ϕ is a test function. Directly sampling from π_{data} may be infeasible, but we assume that we can evaluate its density up to a normalization constant. Instead, we draw samples $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ from a proposal distribution $p(\boldsymbol{x})$ and form the self-normalized estimator

$$\mathbb{E}_{\boldsymbol{x} \sim \pi_{\text{data}}}[\phi(\boldsymbol{x})] \approx \sum_{n=1}^{N} \bar{w}_n\, \phi\big(\boldsymbol{x}^{(n)}\big),$$

where

$$w_n = \frac{\pi_{\text{data}}\big(\boldsymbol{x}^{(n)}\big)}{p\big(\boldsymbol{x}^{(n)}\big)}, \qquad \bar{w}_n = \frac{w_n}{\sum_{m=1}^{N} w_m}.$$

This self-normalized importance sampling (SNIS) estimator is consistent and asymptotically unbiased, with its variance decreasing as N increases.

A.2 Effective Sample Size

The quality of an IS estimator is often measured by its effective sample size (ESS). Two common definitions are:

- *Forward ESS*, which evaluates the overlap with the target distribution:

$$\widehat{\text{ESS}}_{\text{fwd}} = \frac{N^2}{Z^{-1} \sum_{i=1}^{N} w\big(\boldsymbol{x}^{(i)}\big)}, \quad \boldsymbol{x}^{(i)} \sim \pi_{\text{data}}, \tag{11}$$

  where $Z^{-1} \approx \frac{1}{N} \sum_{n=1}^{N} \frac{1}{w_n}$ is the inverse normalizing constant.

- *Reverse ESS*, which measures the concentration of weights under the proposal:

$$\widehat{\text{ESS}}_{\text{rev}} = \frac{\big(\sum_{i=1}^{N} w(\boldsymbol{y}^{(i)})\big)^2}{\sum_{i=1}^{N} w(\boldsymbol{y}^{(i)})^2}, \quad \boldsymbol{y}^{(i)} \sim p. \tag{12}$$

While reverse ESS is widely used for evaluation, it can overestimate performance if p suffers from mode collapse. Forward ESS, by contrast, tends to provide a more reliable measure of overlap between the target and proposal.

A.3 Combining Score-based Diffusion Models with Importance Sampling

We now consider task of estimating the integral $\mathbb{E}_{\boldsymbol{x} \sim \pi_{\text{data}}}[\phi(\boldsymbol{x})]$ where normalized or unnormalized π_{data} is the target density which we can evaluate, and ϕ is the test function of interest. We aim to train a diffusion models to act as the proposal distribution which we can draw samples from. Suppose we have a time interval $[\epsilon, T]$ and we discretize the time interval into N sections with $N{+}1$ time steps such that $\epsilon = t_0 < t_1 < \cdots < t_N = T$.

Assume the conditional proposal distribution of the diffusion model is $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n)$, and the conditional noise distribution is $q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})$ for $n = 1, \ldots, N$. Using SNIS, the integral can be estimated as

$$\mathbb{E}_{\boldsymbol{x} \sim \pi}[\phi(\boldsymbol{x})] \approx \mathbb{E}_{\boldsymbol{x}_0 \sim \pi}[\phi(\boldsymbol{x}_0)] \tag{13}$$

$$= \int \phi(\boldsymbol{x}_0)\pi(\boldsymbol{x}_0)d\boldsymbol{x}_0 \tag{14}$$

$$= \int \phi(\boldsymbol{x}_0)\pi(\boldsymbol{x}_0)\left(\prod_{n=1}^{N} q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})\right) d\boldsymbol{x}_{0:N} \tag{15}$$

$$= \int \phi(\boldsymbol{x}_0)\left(\frac{\pi(\boldsymbol{x}_0)\prod_{n=1}^{N} q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_N)\prod_{n=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n)}\right) p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:N})d\boldsymbol{x}_{0:N} \tag{16}$$

$$\approx \frac{1}{K}\sum_{k=1}^{K} \underbrace{\left(\frac{\pi(\boldsymbol{x}_{t_0}^{(k)})\prod_{n=1}^{N} q(\boldsymbol{x}_{t_n}^{(k)}|\boldsymbol{x}_{t_{n-1}}^{(k)})}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_N^{(k)})\prod_{n=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n-1}}^{(k)}|\boldsymbol{x}_{t_n}^{(k)})}\right)}_{w_k} \phi(\boldsymbol{x}_0) \tag{17}$$

where $\boldsymbol{x}_{0:N}^{(k)}$ for $k = 1, \ldots, K$ are samples from the joint proposal distribution $p_{\boldsymbol{\theta}}$ and w_k for $k = 1, \ldots, K$ are the importance weights. Note that ϵ is chosen such that the error for the first approximation is negligible.

We now need to specify the exact forms of $q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n)$. Recall the forward SDE with zero drift ($\boldsymbol{f} \equiv \boldsymbol{0}$) and diffusion coefficient $g(t) = \sqrt{2t}$:

$$\mathrm{d}\boldsymbol{x}_t = \sqrt{2t}\,\mathrm{d}\boldsymbol{w}_t.$$

Given a time discretization $\epsilon = t_0 < t_1 < \cdots < t_N = T$, the conditional *target* distribution of the forward process is Gaussian:

$$p(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}) = \mathcal{N}\big(\boldsymbol{x}_n; \boldsymbol{x}_{n-1}, (t_n^2 - t_{n-1}^2)\,\boldsymbol{I}\big), \qquad t_n > t_{n-1}. \tag{18}$$

By Bayes' rule, the DDPM-style *reverse* (proposal) distribution conditional on $\boldsymbol{x}_n$ and $\boldsymbol{x}_0$ is

$$p(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_n, \boldsymbol{x}_0) \propto p(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1})\, p(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_0) \tag{19}$$
$$= \mathcal{N}\big(\boldsymbol{x}_{n-1}; \boldsymbol{\mu}(\boldsymbol{x}_n, \boldsymbol{x}_0), \sigma_{\mathrm{ddpm}}^2(n)\,\boldsymbol{I}\big),$$

where

$$\boldsymbol{\mu}(\boldsymbol{x}_n, \boldsymbol{x}_0) = \frac{t_n^2}{t_{n-1}^2}\,\boldsymbol{x}_n + \left(1 - \frac{t_n^2}{t_{n-1}^2}\right)\boldsymbol{x}_0, \tag{20}$$

$$\sigma_{\mathrm{ddpm}}^2(n) = \frac{t_{n-1}^2\left(t_n^2 - t_{n-1}^2\right)}{t_n^2}. \tag{21}$$

In practice, $\boldsymbol{x}_0$ will be predicted via a neural network and thus the reverse conditional is defined as $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n) = p(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n, \hat{\boldsymbol{x}}_0(\boldsymbol{x}_n, t_n; \boldsymbol{\theta}))$. Since both forward and reverse conditionals are Gaussian with closed-form parameters, the joint density ratio over trajectories can be evaluated exactly with minimal overhead.

B VT-DIS for E(3)-Equivariant Diffusion Models

In this appendix we detail how the proposed velocity-tempered diffusion importance sampling (VT-DIS) post-training procedure can be integrated with E(3)-equivariant diffusion models.

Notation. Let $M \geq 2$ denote the number of particles and let n be the spatial dimension. Concatenate the coordinates of all particles into the column vector

$$\boldsymbol{x} = \big(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\big) \in \mathbb{R}^{Mn}, \qquad \boldsymbol{x}^{(i)} \in \mathbb{R}^n.$$

Unless stated otherwise we restrict our attention to the zero–centre-of-mass (CoM) subspace

$$\mathcal{X}_0 = \left\{ \boldsymbol{x} \in \mathbb{R}^{Mn} : \tfrac{1}{M} \sum_{i=1}^{M} \boldsymbol{x}^{(i)} = \boldsymbol{0} \right\}.$$

Change-of-basis matrix. To remove the CoM degree of freedom, we construct a change-of-basis matrix P that maps any $\boldsymbol{x} \in \mathcal{X}_0$ from the Mn-dimensional ambient space to an $(M-1)n$-dimensional Euclidean space. Choose an orthonormal matrix $V_M \in \mathbb{R}^{(M-1) \times M}$ whose rows span the subspace $\{\boldsymbol{z} \in \mathbb{R}^M : \boldsymbol{1}^\top \boldsymbol{z} = 0\}$; one explicit choice is obtained by applying a QR factorization to $I_M - \frac{1}{M}\boldsymbol{1}\boldsymbol{1}^\top$. Define

$$P := V_M \otimes I_n \ \in \ \mathbb{R}^{(M-1)n \times Mn},$$

where $\otimes$ denotes the Kronecker product. For any $\boldsymbol{x} \in \mathcal{X}_0$ we set $\tilde{\boldsymbol{x}} = P\boldsymbol{x} \in \mathbb{R}^{(M-1)n}$. The identities

$$PP^\top = I_{(M-1)n}, \qquad P^\top P = \left(I_M - \tfrac{1}{M}\boldsymbol{1}\boldsymbol{1}^\top\right) \otimes I_n, \tag{22}$$

follow immediately from $V_M V_M^\top = I_{M-1}$ and $V_M^\top V_M = I_M - \frac{1}{M}\boldsymbol{1}\boldsymbol{1}^\top$. The first identity shows that P is an isometry between $\mathcal{X}_0$ and $\mathbb{R}^{(M-1)n}$.

Evaluating and Sampling from $\mathcal{N}_x$ with full covariance For $\boldsymbol{\mu} \in \mathcal{X}_0$ and a positive-definite $\Sigma \in \mathbb{R}^{Mn \times Mn}$ whose image lies in $\mathcal{X}_0$, define the density

$$\mathcal{N}_x(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \mathcal{N}(\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{\mu}}, P\Sigma P^T) \tag{23}$$

$$= \frac{1}{(2\pi)^{\frac{(M-1)n}{2}} \sqrt{\det(P\Sigma P^\top)}} \exp\left[-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top P^\top P \Sigma^{-1} P^\top P(\boldsymbol{x}-\boldsymbol{\mu})\right]. \tag{24}$$

To get samples from this distribution, we have to first sample from the normal distribution with dimension $(M-1)n$ and then we map it back to the Mn dimensional ambient space so that its center of gravity equals zero using the change-of-basis matrix P. This is the general way to get samples for $\mathcal{N}_x$ with full covariance. Things will simplify if the covariance becomes isotropic.

If $\Sigma = \sigma^2 I_{Mn}$ is isotropic, (24) will becomes:

$$\mathcal{N}_x(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) := \frac{1}{(2\pi\sigma^2)^{\frac{(M-1)n}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

because of the fact that $PP^\top = I_{M-1}$. This is the Gaussian distribution constrained on zero CoM subspace with isotropic covariance (Hoogeboom et al., 2022; Xu et al., 2022). To sample from this distribution, we could sample in the Mn dimensional ambient space directly, and subtract its center of mass (Xu et al., 2022; Satorras et al., 2021).

B.1 Proposed Covariance Structure Satisfies E(3)-Equivariance

A diffusion model is E(3)-equivariant if, for every orthogonal transformation $R \in \mathrm{O}(Mn)$,[3] (i) the forward-process endpoint distribution $p(\boldsymbol{x}_T)$ is invariant, and (ii) the reverse-process conditionals $p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ are equivariant:

$$p(R\boldsymbol{x}_{t-1} \mid R\boldsymbol{x}_t) = p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \quad \forall t \in \{1, \ldots, T\}.$$

Invariance of the isotropic Gaussian prior $p(\boldsymbol{x}_T) = \mathcal{N}_x(\boldsymbol{0}, T^2 I_{Mn})$ is immediate. We therefore focus on the conditional law.

Lemma 1. *Let the reverse transition be*

$$p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}_x(\boldsymbol{x}_{t-1}; \boldsymbol{x}_t, \Sigma), \qquad \Sigma = B \otimes I_n,$$

where $B \in \mathbb{R}^{M \times M}$ is positive definite. Then the reverse kernel is E(3)-equivariant.

---

[3] Translations are excluded because we restrict to the zero–CoM subspace $\mathcal{X}_0$; cf. Sec. B.

Proof. Every spatial rotation acts identically on all M particles and can therefore be written as the block-orthogonal matrix

$$R = I_M \otimes R_n, \qquad R_n \in \mathrm{O}(n).$$

Recall that the reverse kernel is defined on the $(M-1)n$-dimensional coordinate $\tilde{\boldsymbol{x}} := P\boldsymbol{x}$ via

$$p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}\big(P\boldsymbol{x}_{t-1}; P\boldsymbol{x}_t, P\Sigma P^T\big), \qquad \Sigma = B \otimes I_n, \quad B \succ 0.$$

For arbitrary $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}_0$ we have

$$
\begin{aligned}
\log p(R\boldsymbol{y} \mid R\boldsymbol{x}) &= \log \mathcal{N}(PR\boldsymbol{y}; PR\boldsymbol{x}, P\Sigma P^\top) \\
&= -\tfrac{1}{2}\big(PR\boldsymbol{y} - PR\boldsymbol{x}\big)^\top P\Sigma^{-1}P^\top\big(PR\boldsymbol{y} - PR\boldsymbol{x}\big) + \mathrm{const} \\
&= -\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\top R^\top P^\top P\Sigma^{-1}P^\top PR(\boldsymbol{y} - \boldsymbol{x}) \\
&\overset{(a)}{=} -\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\top P^\top PR^\top \Sigma^{-1}RP^\top P(\boldsymbol{y} - \boldsymbol{x}) \\
&\overset{(b)}{=} -\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\top P^\top P\Sigma^{-1}P^\top P(\boldsymbol{y} - \boldsymbol{x}) \\
&= -\tfrac{1}{2}\big(P\boldsymbol{y} - P\boldsymbol{x}\big)^\top P\Sigma^{-1}P^\top\big(P\boldsymbol{y} - P\boldsymbol{x}\big) \\
&= \log \mathcal{N}(P\boldsymbol{y}; P\boldsymbol{x}, P\Sigma P^\top) \\
&= \log p(\boldsymbol{y} \mid \boldsymbol{x}),
\end{aligned}
$$

where step (a) uses the fact that $P^\top P$ commutes with R:

$$RP^\top P = (I_n \otimes R_n)(V_M \otimes I_n) = V_M \otimes R_n = P^\top PR$$

and step (b) uses the fact that

$$
\begin{aligned}
R^\top \Sigma^{-1}R &= (R^\top \Sigma R)^{-1} \\
&= ((I_M \otimes R_n^\top)(B \otimes I_n)(I_M \otimes R_n))^{-1} \\
&= (B \otimes R_n^\top R_n)^{-1} \\
&= (B \otimes I_n)^{-1} \\
&= \Sigma^{-1}
\end{aligned}
$$

Hence we have proved that our proposed covariance form satisfies the E(3)-equivariance required. $\qquad\square$

B.2 Permutation Equivariance for Many-Particle Systems

Datasets such as DW-4 and LJ-13 treat all particles as indistinguishable; hence the reverse kernel must be equivariant under any permutation of the M particles.

Lemma 2 (Permutation equivariance)**.** *Let the covariance in the reduced coordinate space be*

$$\Sigma = B \otimes I_n, \qquad B = (b-a)I_{M-1} + a\,\mathbf{1}\mathbf{1}^\top,$$

with parameters $b > 0$ and $a \in \mathbb{R}$ chosen such that $b - a > 0$ and $b + (M-2)a > 0$, ensuring $B \succ 0$. For any permutation matrix $S_M \in \mathbb{R}^{M \times M}$ and the block permutation $S := S_M \otimes I_n$, the reverse transition satisfies $p(S\boldsymbol{y} \mid S\boldsymbol{x}) = p(\boldsymbol{y} \mid \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}_0$.

Proof. **(i) Commutation with the CoM projector.** Using $P^\top P = (V_M \otimes I_n)$ (Sec. B) and the fact that $S_M \mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top S_M = \mathbf{1}^\top$,

$$SP^\top P = \big(S_M\big(I_M - \tfrac{1}{M}\mathbf{1}\mathbf{1}^\top\big)\big) \otimes I_n = \big(S_M - \tfrac{1}{M}\mathbf{1}\mathbf{1}^\top\big) \otimes I_n = P^\top PS,$$

so S commutes with the projector onto the zero–CoM subspace.

(ii) Invariance of the covariance. Because S_M merely reorders rows and columns,

$$S_M B S_M^\top = (b-a)\, I_{M-1} + a\, S_M \mathbf{1}\mathbf{1}^\top S_M^\top = (b-a)\, I_{M-1} + a\, \mathbf{1}\mathbf{1}^\top = B,$$

whence $S\Sigma S^\top = \Sigma$.

(iii) Equality of densities. With the same quadratic-form argument used in the E(3)-equivariance proof,

$$(\boldsymbol{y} - \boldsymbol{x})^\top P^\top P\, \Sigma^{-1}\, P^\top P(\boldsymbol{y} - \boldsymbol{x}) \;=\; (S\boldsymbol{y} - S\boldsymbol{x})^\top P^\top P\, \Sigma^{-1}\, P^\top P(S\boldsymbol{y} - S\boldsymbol{x}),$$

establishing $p(S\boldsymbol{y} \mid S\boldsymbol{x}) = p(\boldsymbol{y} \mid \boldsymbol{x})$. $\qquad\square$

B.3 Permutation Equivariance for General Molecular Distributions

Again, we aim to show that our choice of B_ϕ satisfies the permutation-equivariance constraint under the parametrization proposed in Section 4.3.

Permutation equivariance. Let S_M denote any permutation matrix that reorders atoms *within* each label class (i.e., $L_{S_M(i)} = L_i$). Analogous to the proof in Section B.2, because any permutation matrix commutes with the CoM projector, this condition is automatically satisfied. All that remains is to prove the invariance of the covariance. Because the two parameterizations above depend on the indices only through the labels, we have

$$S_M\, B_\phi(n)\, S_M^\top = B_\phi(n).$$

Consequently, the lifted covariance

$$\Sigma_\phi(n) = B_\phi(n) \otimes I_n$$

satisfies

$$S\, \Sigma_\phi(n)\, S^\top = \Sigma_\phi(n)$$

for every block permutation $S = S_M \otimes I_n$, so the reverse kernel remains permutation-equivariant by the argument of the above section.

C Implementation Details

All experiments were run on a single NVIDIA A100 GPU (80 GB); hence every "GPU hour" refers to wall-clock hours on that device.

C.1 Implementation of the objective function

Recall the objective (the α-divergence with $\alpha = 2$):

$$\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_q\left[\frac{q(\mathbf{x}_{0:N})}{p_\phi(\mathbf{x}_{0:N})}\right] \approx \frac{1}{M}\sum_{m=1}^{M} \frac{q(\mathbf{x}_{0:N}^{(m)})}{p_\phi(\mathbf{x}_{0:N}^{(m)})} = \frac{1}{M}\sum_{m=1}^{M} w^{(m)},$$

where $\mathbf{x}_{0:N}^{(m)} \sim q$, $w^{(m)} := \frac{q(\mathbf{x}_{0:N}^{(m)})}{p_\phi(\mathbf{x}_{0:N}^{(m)})}$, and M is the number of samples used to estimate the objective.

In practice, we optimize the *log* of the α-divergence for numerical stability, which can be computed efficiently using the log-importance weights via the `logsumexp` trick. Specifically,

$$\mathcal{J}(\boldsymbol{\phi}) := \log \mathcal{L}(\boldsymbol{\phi}) \approx \log\left(\frac{1}{M}\sum_{m=1}^{M} \exp(\log w^{(m)})\right) = \texttt{logsumexp}\big(\{\log w^{(m)}\}_{m=1}^{M}\big) - \log M. \qquad (25)$$

Hence, we work entirely in log space, which yields a numerically stable estimator of the loss.

Table 2: Score-model training time on a single A100.

Dataset	Architecture	Layers	Hidden dim	Batch Size	Iterations	GPU-h
GMM-2 (d=50)	MLP	7	512	5000	100 000	0.6
GMM-2 (d=100)	MLP	7	512	5000	100 000	1.0
DW-4	EGNN	4	128	512	100 000	0.5
LJ-13	EGNN	6	128	512	100 000	8.0
Alanine dipeptide	EGNN	7	256	512	100 000	8.0

C.2 Datasets

- **GMM-2.** A two-component Gaussian mixture used as a toy benchmark. Training and evaluation samples are drawn directly from the target distribution.

- **DW-4 and LJ-13.** We use the 10^7 configurations released by Klein et al. (2023). A subset of 10^5 configurations trains the score model and VT-DIS; the remainder forms the test set.

- **Alanine dipeptide.** We adopt the molecular dataset of Midgley et al. (2023), which provides 10^5 training samples and 10^6 test samples.

C.3 Training the Score Model

Our training pipeline—including parameterization and preconditioning—follows Karras et al. (2022).

- **GMM-2.** A seven-layer MLP (hidden width 512), identical to the network used by Akhound-Sadegh et al. (2024) for the GMM-40 task. Training runs for 100,000 iterations with a batch size of 1,024 for both $d = 50$ and $d = 100$.

- **DW-4.** EGNN backbone with 4 layers and a hidden width of 128.

- **LJ-13.** EGNN backbone with 6 layers and a hidden width of 128.

- **Alanine Dipeptide.** EGNN backbone with 7 layers and a hidden width of 256; node features include atom types and bond information. Scaling atomic coordinates by 10 during training improves sample quality; network outputs are divided by 10 at inference.

All models are trained for 100,000 iterations with a batch size of 512 using the Adam optimizer, a learning rate of 0.001, and a cosine annealing schedule with a minimum learning rate of 1×10^{-6}. Table 2 summarizes these details.

C.4 VT-DIS Post-training

We follow Algorithm 1 exactly. With the score network frozen, we optimize the per-step covariance matrices by minimizing the α-divergence ($\alpha = 2$) between the forward and reverse processes using the same training set as for score learning. Each run comprises 5,000 optimization steps with a batch size of 512 (256 for alanine dipeptide). All runs use the Adam optimizer with a learning rate of 0.01 and a cosine annealing schedule with a minimum learning rate of 1×10^{-6}. Table 3 summarizes these settings.

The reported training time corresponds to tuning the isotropic covariance; diagonal and full covariance forms incur similar runtimes. We fix the number of time steps to 100; empirically, with fixed epochs and batch sizes, training time increases linearly with the number of sampling steps.

Table 3: VT-DIS post-training time on a single A100.

Dataset	Batch Size	Iterations	Steps K	GPU-h
GMM-2 ($d = 50$)	512	5 000	100	1.8
GMM-2 ($d = 100$)	512	5 000	100	3.5
DW-4	512	5 000	100	0.9
LJ-13	512	5 000	100	2.7
Alanine dipeptide	128	5 000	100	4.3

Table 4: VT-DIS sampling time on a single A100.

Dataset	Batch Size	Steps K	GPU-min
GMM-2 ($d = 50$)	100,000	100	0.25
GMM-2 ($d = 100$)	100,000	100	0.5
DW-4	100,000	100	1.3
LJ-13	100,000	100	4.3
Alanine dipeptide	10,000	100	3

C.5 Evaluation Metrics

We measure efficiency using forward and reverse effective sample sizes (ESS), defined in Eqs. (11) and (12). The number of Monte Carlo draws for each estimate is specified in the corresponding figure or table caption. Table 4 summarizes the sampling time for each target distribution. Empirically, sampling time increases linearly with the number of sampling steps.

D Additional Experimental Results

D.1 Computational Cost for Solving Probability Flow ODE

In this section, we compare effective sample size (ESS) and inference-time computational cost for an additional baseline: PF-ODE with JVP-based trace estimation. Table 5 summarizes the results. Using Hutchinson's trace estimator reduces computational cost; however, with a single probe vector the ESS drops sharply, whereas using 10 probe vectors increases both cost and ESS. VT-DIS retains the lowest inference-time cost among all baselines considered. We emphasize that although Hutchinson's estimator is unbiased for the instantaneous divergence, using stochastic divergence estimates to compute importance weights can introduce bias in self-normalized IS estimates. Consistent with prior work on these datasets (e.g., (Klein et al., 2023; Tan et al., 2025)), our main-text comparison computes exact divergences when applying IS to avoid this bias. Accordingly, we caution against using stochastic trace estimators for IS and explicitly flag this caveat in the PF-ODE+trace results.

D.2 Optimized Covariance Matrices

Figure 6 contrasts the per–step variance–scaling factor of VT-DIS, η_n (see (6)), with the unit baseline of the original DDPM. Although we display the isotropic case, the same qualitative pattern appears for diagonal and low–rank covariances.

Three regimes. Reading the t-axis from left ($t \approx 10^{-3}$) to right ($t \approx 10^2$) follows the reverse SDE from data space to pure noise. The learned profile is *not* monotone but exhibits three characteristic regimes:

1. **Data–dominated regime** ($t \lesssim 10^{-1}$). Here the signal-to-noise ratio (SNR) is high, the score network is the main source of model error, and under-dispersion relative to the true forward kernel results in heavy

Table 5: Performance of importance sampling with diffusion on DW-4, LJ-13 and ALDP with 10^4 samples on a single A100 GPU. The number in parentheses after each dataset denotes the sampling steps. Methods: reverse SDE (VT-DIS) or PF ODE (exact or Hutchinson with n probe vectors). ↑ = higher is better, ↓ = lower is better.

Dataset (steps)	Method	Forward ESS (%) ↑	Reverse ESS (%) ↑	GFLOP / sample ↓
DW-4 (100)	SDE (Ours)	44.32	46.76	0.81
	ODE (Exact)	94.29	83.60	7.70
	ODE (Est. $n=1$)[†]	37.70	0.95	1.56
	ODE (Est. $n=10$)[†]	80.60	82.67	8.27
LJ-13 (200)	SDE (Ours)	15.87	17.69	26.20
	ODE (Exact)	79.21	77.42	2173.00
	ODE (Est. $n=1$)[†]	12.35	9.50	52.26
	ODE (Est. $n=10$)[†]	61.12	66.10	285.62
ALDP (200)	SDE (Ours)	0.00	1.05	33.70
	ODE (Exact)	0.00	35.39	45 967.00
	ODE (Est. $n=1$)[†]	0.00	0.74	702.78
	ODE (Est. $n=10$)[†]	0.00	22.31	3850.39

[†] Hutchinson trace estimator with n probe vectors; introduces bias in importance sampling and can misstate ESS.

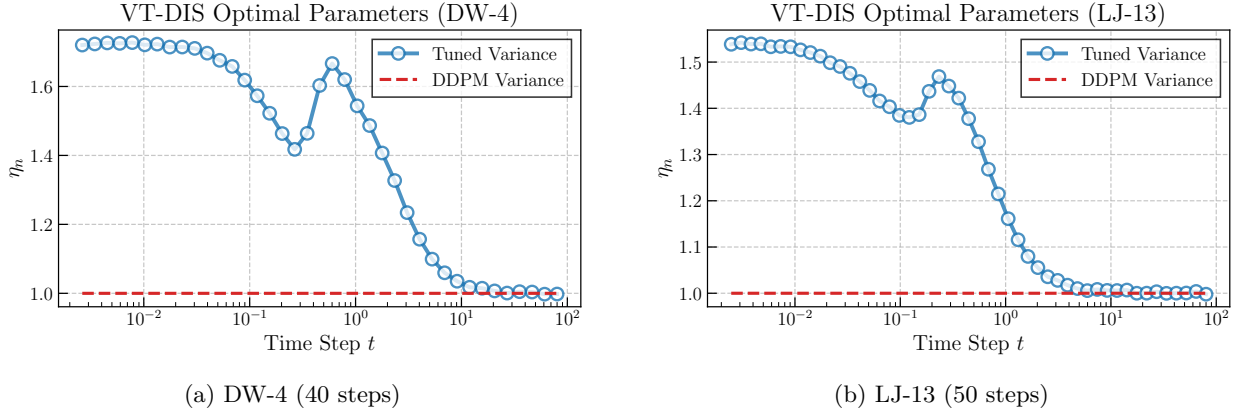(a) DW-4 (40 steps)   (b) LJ-13 (50 steps)

Figure 6: **Optimal variance scaling η_n.** VT-DIS inflates early-time variance to improve coverage, adapts non-monotonically across the transition regime, and converges to the DDPM value ($\eta_n = 1$, red dashed line) in the pure-noise limit.

importance-weight tails. VT-DIS compensates by *inflating* the variance ($\eta_n \approx 1.6$–1.7), which widens the proposal and improves coverage.

2. **Transition regime** ($10^{-1} \lesssim t \lesssim 10^1$). As the SNR drops, the target conditional gradually morphs from highly non-Gaussian to approximately Gaussian. The mismatch between the analytic DDPM variance (optimal for the *Gaussian* limit) and the true kernel therefore decreases. VT-DIS first reduces the scaling (a local minimum at $t \approx 0.3$), then overshoots to a peak at $t \approx 1$, before descending. The intermediate bump arises where the learned score D_θ tends to *over-estimate* the gradient magnitude (a common artefact of score matching when the SNR is intermediate, e.g. $\mathcal{O}(1)$) (Xu et al., 2023). Increasing the proposal variance temporarily counter-balances this over-confidence, reducing the second-moment of the importance weights.

3. **Noise regime** ($t \gtrsim 10^1$). The state distribution is now close to $\mathcal{N}(0, t^2 I)$, for which the DDPM variance is *exactly* optimal. Accordingly, η_n relaxes to 1, confirming that VT-DIS does not perturb steps that are already statistically efficient.
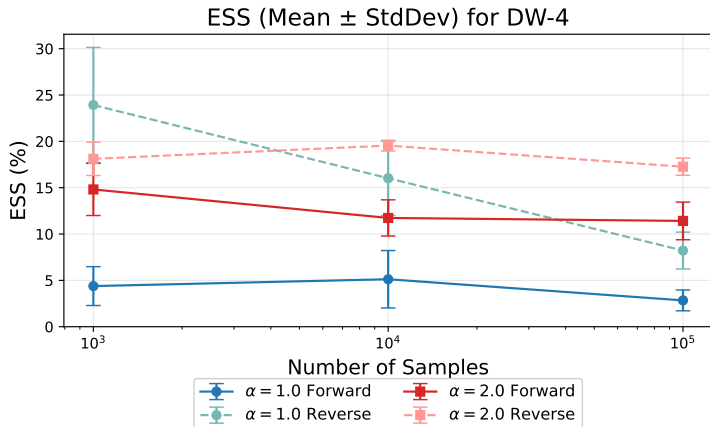
Figure 7: **KL ($\alpha = 1$) vs. $\alpha = 2$ divergence on DW-4.** $\alpha = 2$ tuning (red) attains higher, size-independent ESS, whereas KL tuning (blue) collapses for large M.

Implications. The variance inflations at early steps are crucial: without them, trajectories started near the data manifold place negligible mass in regions where the forward kernel has significant probability, causing weight explosion and the near-zero forward ESS observed for the vanilla baseline. Conversely, keeping $\eta_n \to 1$ in the noise regime ensures that VT-DIS retains the favourable properties of the original DDPM sampler at large time steps, adding no unnecessary dispersion..

D.3 KL versus α-divergence

The VT-DIS objective in (10) minimises the α-divergence with $\alpha = 2$, which is proportional to the *variance* of importance weights and therefore maximizes ESS directly (Midgley et al., 2022). For completeness we retrain the covariance parameters on DW-4 using the forward KL divergence ($\alpha = 1$) and compare the resulting samplers.

Figure 7 plots forward (solid) and reverse (dashed) ESS as functions of the sample size M. Tuning with $\alpha = 2$ yields consistently higher and *stable* ESS in both directions. In contrast, KL-tuned parameters give (i) low forward ESS and (ii) reverse ESS that *decreases* with M, a classic symptom of heavy weight tails caused by poor proposal coverage. These observations confirm that controlling the second moment of weights ($\alpha = 2$) is crucial for robust importance sampling, whereas KL divergence alone fails to regularise the tails.

E Broader Impacts

Our method enhances the efficiency and accessibility of unbiased molecular sampling, potentially speeding up the discovery of catalysts, drugs, and new materials while reducing reliance on costly experiments. At the same time, stronger samplers could be misused to design toxic or environmentally harmful compounds more rapidly. There is also a risk that institutions with advanced computing resources will widen the gap with those lacking such infrastructure, reinforcing inequities in scientific research.