Rao-Blackwell Gradient Estimators for Equivariant Denoising Diffusion

Vinh Tong[1,2]*, **Trung-Dung Hoang**[4]*, **Anji Liu**[5], **Guy Van den Broeck**[3], **Mathias Niepert**[1,2]
[1]University of Stuttgart, [2]IMPRS-IS, [3]UCLA, [4]University of Bern, [5]National University of Singapore
vinh.tong@ki.uni-stuttgart.de

Abstract

In domains such as molecular and protein generation, physical systems exhibit inherent symmetries that are critical to model. Two main strategies have emerged for learning invariant distributions: designing equivariant network architectures and using data augmentation to approximate equivariance. While equivariant architectures preserve symmetry by design, they often involve greater complexity and pose optimization challenges. Data augmentation, on the other hand, offers flexibility but may fall short in fully capturing symmetries. Our framework enhances both approaches by reducing training variance and providing a provably lower-variance gradient estimator. We achieve this by interpreting data augmentation as a Monte Carlo estimator of the training gradient and applying Rao–Blackwellization. This leads to more stable optimization, faster convergence, and reduced variance, all while requiring only a single forward and backward pass per sample. We also present a practical implementation of this estimator—incorporating the loss and sampling procedure—through a method we call *Orbit Diffusion*. Theoretically, we guarantee that our loss admits equivariant minimizers. Empirically, Orbit Diffusion achieves state-of-the-art results on GEOM-QM9 for molecular conformation generation, improves crystal structure prediction, and advances text-guided crystal generation on the Perov-5 and MP-20 benchmarks. Additionally, it enhances protein designability in protein structure generation. Code is available at `https://github.com/vinhsuhi/Orbit-Diffusion.git`.

1 Introduction

Diffusion models have emerged as powerful methods for modeling complex distributions (Ho et al., 2020; Song et al., 2021a,b; Karras et al., 2022), with applications in domains such as molecular and protein generation (Zhang et al., 2023; Vignac et al., 2023; Anand & Achim, 2022). Many physical systems, such as molecules or crystals, exhibit inherent symmetries. For example, a molecule's physical properties remain unchanged under rotations in 3D space (Hoogeboom et al., 2022; Jing et al., 2022). Modeling such data requires learning distributions that are invariant under the action of a group G. This setting is naturally captured by the notion of G-*invariant* distribution $q(x_0)$ that are invariant under transformations from G (Chen et al., 2024; Köhler et al., 2020).

Two main strategies have emerged for learning G-*invariant* distributions: (1) designing equivariant network architectures, and (2) using data augmentation to approximate equivariance. Equivariant architectures, such as equivariant denoisers, ensure symmetry by construction (Hoogeboom et al., 2022; Klein et al., 2024; Igashov et al., 2024), but are often less efficient due to increased architectural complexity and can pose optimization challenges (Brehmer et al., 2024; Abbe & Boix-Adserà, 2022). In contrast, data augmentation is a flexible and widely used alternative that approximates equivariant training by sampling transformed versions of the input. While this approach scales easily and has

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

become increasingly popular in the community, especially for large-scale models (Abramson et al., 2024b; Geffner et al., 2025), its effectiveness in capturing symmetries may vary depending on the group and application domain. In some settings, such as molecular dynamics and structural biology, explicit equivariance remains beneficial (Anand & Achim, 2022; Batatia et al., 2022; Zaverkin et al., 2024). In this work, we propose a framework that provably improves *both* approaches: we introduce a novel form of implicit data augmentation by computing the denoising target as a weighted average over group orbits, which reduces variance and improves the training of equivariant denoisers.

We revisit data augmentation from a principled perspective and interpret it as a Monte Carlo estimator of the gradient of a symmetrized loss. This loss is defined over a fully symmetrized dataset, which yields an empirical distribution invariant under G (Chen et al., 2024). While traditional data augmentation uses one or few samples from this augmented dataset, we instead apply Rao–Blackwellization to derive a new estimator with provably lower variance. Specifically, we decompose the gradient as an outer expectation over data and an inner conditional expectation over group actions. Replacing the noisy sample-based target with its conditional expectation yields a lower-variance gradient estimator, while preserving equivariance.

To translate our theoretical insights into a practical method, we develop an efficient implementation of the proposed gradient estimator. This approach integrates the symmetrized loss and variance reduction into a modified training objective, without increasing the computational cost. Our implementation requires only a single forward and backward pass per sample, and it is compatible with both equivariant and non-equivariant architectures. We refer to this practical method as *Orbit Diffusion*. By implicitly applying Rao–Blackwellization through a tailored loss formulation and sampling scheme, Orbit Diffusion enables stable optimization and improved generalization across a wide range of symmetry groups and tasks.

We provide theoretical guarantees that our symmetrized loss admits equivariant minimizers and that our gradient estimator has strictly smaller variance compared to existing methods. Empirically, we demonstrate strong performance across multiple domains. Our method achieves state-of-the-art results on GEOM-QM9 for molecular conformation generation, enhances crystal structure prediction, and improves text-guided crystal generation on the Perov-5 and MP-20 benchmarks. Moreover, our approach is compatible with non-equivariant denoisers; in particular, it improves the designability of protein structures generated by PROTEINA (Geffner et al., 2025).

2 Background

Groups. A *group* is a mathematical structure comprising a set G and a binary operation $m : G \times G \to G$ that combines two elements of G. A *group action* $g \in G$ defines how the group G acts on a set Ω, such as a set of geometric objects.[2] We restrict our attention to *locally compact isometry groups*. Locally compact isometry groups encompass a broad class of transformation groups that preserve distances and possess a well-behaved topological structure. Examples include the permutation group S_n, the orthogonal group $\mathsf{O}(d)$, and the special orthogonal group $\mathsf{SO}(d)$.

Invariance and Equivariance. A function $f : \Omega \to \mathbb{R}$ is said to be *G-invariant* if for all $g \in G$ and $x \in \Omega$, it satisfies $f(g \circ x) = f(x)$. This means the function value does not change under the action of any group element. A function $f : \Omega \to \Omega$ is said to be *G-equivariant* if for all $x \in \Omega$, f commutes with any group action $g \in G$: $f(g \circ x) = g \circ f(x)$.

Invariant and Equivariant Distributions. A probability distribution $p(x)$ defined on a set Ω is said to be *G-invariant* under the action of a group G if the probability of any measurable subset $A \subseteq \Omega$ remains unchanged under the transformation induced by any group action $g \in G : p(g \circ x \in A) = p(x \in A)$. A conditional distribution $p(y \mid x)$, where $x \in \Omega$ and $y \in \Omega$, is said to be *G-equivariant* if for all $g \in G$, the following condition holds: $p(g \circ y \mid g \circ x) = p(y \mid x)$..

Group Symmetrization. Let S_G be the symmetrization operator under group G, transforming any distribution $p(x)$ into a *G-invariant*-invariant distribution, denoted as the *G-symmetrized* distribution:

$$S_G[p](x) := \int_G p(g \circ x)\, \mathrm{d}\mu_G(g), \qquad \text{where } \mu_G \text{ is the Haar measure on } G. \tag{1}$$

[2]When the group acts on a vector space V, we do not distinguish between the abstract group element $g \in G$ and its linear representation $\rho(g) : G \to \mathrm{GL}(V)$. For simplicity, we write $g \circ x$ to denote the action $\rho(g)(x)$.

Diffusion Models and Equivariant Diffusion Models. Diffusion models (Ho et al., 2020) are a class of generative models that construct complex data distributions by iteratively transforming simple noise distributions through a learned denoising process. Formally, given data $x_0 \sim q(x_0)$, the forward process generates a sequence x_t over time $t \in [0, T]$ using a stochastic differential equation (SDE) (Song et al., 2021b) or a discrete Markov chain (Ho et al., 2020), such as:

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where α_t and σ_t are a time-dependent scaling factor and a noise factor that determines the level of noise added at each step, respectively. The noise should be increasingly added to the sample so that at time $t = T$, $q(x_T) \approx \mathcal{N}(0, I)$.

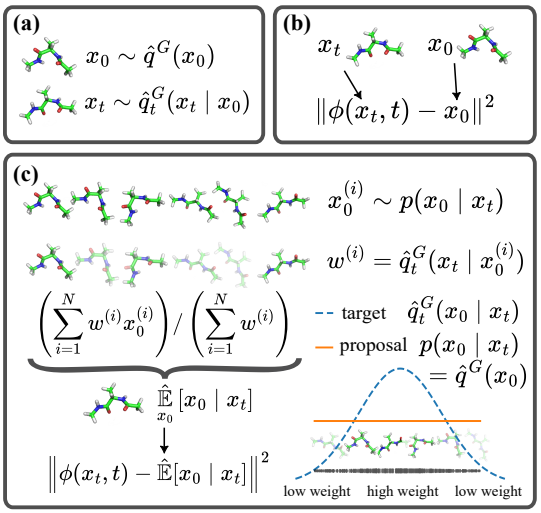
The reverse process, parameterized by a neural network $\phi_\theta(x_t, t)$, approximates a clean sample x_0 given its noisy version x_t. The training objective typically involves minimizing a reweighted form of the denoising loss (Ho et al., 2020; Song et al., 2021b; Karras et al., 2022):

$$\mathcal{L} = \mathbb{E}_{t\sim\mathcal{U}(0,T),(x_0,x_t)\sim q(x_0,x_t)} \left[\omega(t)\|\phi_\theta(x_t, t) - x_0\|^2\right], \tag{2}$$

where $\omega(t)$ is a time-dependent loss weight. For notational simplicity, we omit this term throughout the remainder of the paper. To sample from the diffusion model, we begin with a noise vector $x_T \sim \mathcal{N}(0, I)$ and iteratively apply the learned reverse process to transform it into a data sample x_0 using the trained model ϕ_θ. The reverse process can involve solving an ODE or SDE numerically (Song et al., 2021b; Karras et al., 2022; Lu et al., 2022; Tong et al., 2025).

Equivariant diffusion models extend standard diffusion by enforcing equivariance of the neural network denoiser. Specifically, the denoiser ϕ_θ is said to be G-*equivariant* if it satisfies $\phi_\theta(g \circ x_t, t) = g \circ \phi_\theta(x_t, t)$ for all $g \in G$.

3 Method



Algorithm 1: Orbit Diffusion with RB.

Sample a data point $x_0 \sim \hat{q}^G(x_0)$
Sample a noise level t
Generate a noisy sample $x_t \sim \hat{q}_t^G(x_t \mid x_0)$
for $i = 1$ to N **do**
  Sample a group element $g^{(i)} \sim \nu_t(g)$
  Compute orbit sample $x_0^{(i)} = g^{(i)} \circ x_0$
  Compute $w^{(i)} = \hat{q}_t^G(x_t \mid x_0^{(i)}) / \nu_t(g^{(i)})$
end for
Approximate $\mathbb{E}[x_0 \mid x_t]$ with SNIS:

$$\hat{\mathbb{E}}[x_0 \mid x_t] = \left(\sum_{i=1}^N w^{(i)} x_0^{(i)}\right) / \left(\sum_{i=1}^N w^{(i)}\right)$$

Backpropagate gradients of the loss:

$$\left\|\phi(x_t, t) - \hat{\mathbb{E}}[x_0 \mid x_t]\right\|^2$$

Figure 1: Gradient estimation strategies for training (approximately) equivariant diffusion models: (a) Sampling from the symmetrized joint distribution to obtain x_0 and x_t. (b) The standard data augmentation approach, which directly uses these samples for training. (c) The proposed method, leveraging self-normalizing importance sampling (SNIS) to estimate the inner conditional expectation. Both (b) and (c) require a single neural function evaluation per gradient step, but (c) has lower variance than (b). The pseudo-code for the Rao-Blackwell estimator with SNIS is shown on the right.

Let G be a symmetry group, such as the group of Euclidean rotations. Our goal is to learn a G-invariant data-generating distribution $q(x_0)$. However, the observed distribution $\hat{q}$ (from which we obtain training samples) is generally not G-invariant due to dataset biases. For example, in molecular datasets, each molecule may be stored in a canonical but arbitrary orientation, even though physically

all rotated versions are equally probable under q. As a result, training directly on $\hat{q}$ would lead to a model that may not respect the underlying symmetry.

This issue is widely recognized in the literature, and two standard solutions are:

1. **Equivariant model:** Train with the diffusion loss in Equation (2) using a G-equivariant network $\phi(x_t, t)$.
2. **Data augmentation:** Train a non-equivariant ϕ with Equation (2), augmenting data with random group actions to encourage approximate G-invariance.

The second approach has been widely adopted—several high-profile non-equivariant models achieve strong results through augmentation (Abramson et al., 2024b; Geffner et al., 2025). However, empirical evidence shows that augmentation offers *no benefit* for already equivariant models, a result we formally prove in Appendix B.2.

3.1 From Symmetrized Loss to High-Variance Gradient Estimators

To unify these approaches, consider the *symmetrized data distribution* from Equation (1), and the forward noising kernel $\hat{q}_t^G(x_t \mid x_0)$ (e.g., a Gaussian in denoising diffusion) with marginal $\hat{q}_t^G(x_t)$. The *symmetrized diffusion loss* at time t is

$$\mathcal{L}_t^G(\phi) = \mathbb{E}_{x_0 \sim \hat{q}^G} \mathbb{E}_{x_t \sim \hat{q}_t^G(\cdot|x_0)} \left[\|\phi(x_t, t) - x_0\|^2 \right]. \tag{3}$$

Both the equivariant model and the data augmentation approach can be viewed as implicitly minimizing $\mathcal{L}_t^G(\phi)$ (Chen et al., 2024). The *true* gradient of this loss is

$$\nabla_\phi \mathcal{L}_t^G = \mathbb{E}_{x_t \sim \hat{q}_t^G} \mathbb{E}_{x_0 \sim \hat{q}_t^G(\cdot|x_t)} \left[2\big(\phi(x_t, t) - x_0\big) \right], \tag{4}$$

where the expectations are taken over the full symmetrized joint distribution.

In practice, we do not have access to the exact expectations in Equation (4). Instead, we construct a Monte Carlo *gradient estimator* by sampling $x_0 \sim \hat{q}$ (or $\hat{q}^G$ in the augmented/equivariant case), then sampling $x_t \sim \hat{q}_t^G(\cdot \mid x_0)$, and using the single-sample estimate $\widehat{\nabla_\phi \mathcal{L}_t^G} = 2\big(\phi(x_t, t) - x_0\big)$. This estimator is *unbiased*, but, as in other diffusion training setups, it can exhibit *high variance* (Kingma et al., 2021; Xu et al., 2023; Nichol & Dhariwal, 2021). Increasing the batch size reduces variance but requires more forward passes of the neural network ϕ, raising computational cost.

In this work, we introduce a class of Rao–Blackwellized gradient estimators that *provably reduce variance* while remaining unbiased, and can be applied to both the equivariant and augmentation-based training strategies.

3.2 Rao–Blackwellized Gradient Estimator

Our *key observation* is that $\phi(x_t, t)$ does not depend on x_0. This allows us to move it outside the inner expectation in Equation (4), yielding

$$\nabla_\phi \mathcal{L}_t^G = \mathbb{E}_{x_t \sim \hat{q}_t^G} \Big[2\big(\phi(x_t, t) - \underbrace{\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0|x_t)}[x_0]}_{\mathbb{E}[x_0|x_t]} \big) \Big]. \tag{5}$$

Replacing x_0 with its conditional mean $\mathbb{E}[x_0 \mid x_t]$ yields a *Rao–Blackwellized (RB) gradient estimator*, which remains unbiased and has variance no greater than the original—strictly less unless $x_0 \mid x_t$ is deterministic, a situation that rarely occurs in generative modeling where x_0 is typically stochastic given x_t. This improvement requires no additional neural network evaluations, only a more accurate target. To avoid custom backward passes, we can minimize:

$$\mathcal{L}_t^{\mathrm{RB}}(\phi) = \mathbb{E}_{x_t \sim \hat{q}_t^G} \left[\|\phi(x_t, t) - \mathbb{E}[x_0 \mid x_t]\|^2 \right] \tag{6}$$

Variance reduction guarantee

Theorem 1. *Let $\widehat{\nabla}_\phi$ be the Monte Carlo gradient from Equation* (4) *and $\widehat{\nabla}_\phi^{(RB)}$ be from Equation* (5)*. If $\mathbb{E}[x_0 \mid x_t]$ can be computed exactly, then*

$$\mathrm{Var}\big(\widehat{\nabla}_\phi^{(RB)}\big) \leq \mathrm{Var}\big(\widehat{\nabla}_\phi\big),$$

with strict inequality unless $x_0 \mid x_t$ is a Dirac delta.

The challenge now is estimating the loss target $\mathbb{E}[x_0 \mid x_t]$ accurately and efficiently. We address this next using self-normalized importance sampling (SNIS).

3.3 Estimating the Conditional Expectation

A central challenge in computing the gradient estimator is evaluating the conditional expectation $\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0|x_t)}[x_0]$, which is generally intractable:

$$\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0|x_t)}[x_0] = \int_\Omega x_0 \, \hat{q}_t^G(x_0 \mid x_t) \, \mathrm{d}x_0. \tag{7}$$

This expectation can be approximated by drawing independent samples $x_0^{(1)}, \ldots, x_0^{(N)} \sim \hat{q}_t^G(x_0 \mid x_t)$ and computing the sample mean. Unfortunately, we cannot sample efficiently and directly from $\hat{q}_t^G(x_0 \mid x_t)$. Using Bayes' rule:

$$\hat{q}_t^G(x_0 \mid x_t) \propto \hat{q}_t^G(x_t \mid x_0) \hat{q}^G(x_0), \tag{8}$$

where $\hat{q}_t^G(x_t \mid x_0)$ is available in closed form, but $\hat{q}^G(x_0)$ is intractable due to integration over the group orbit of x_0.

To address this, we use self-normalized importance sampling (SNIS) with a proposal distribution $p(x_0 \mid x_t)$ that shares the same intractable orbit-integral structure as $\hat{q}^G(x_0)$, allowing cancellation of the problematic terms in the importance weights. The conditional expectation is approximated as:

$$\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0|x_t)}[x_0] \approx \frac{\sum_{i=1}^N x_0^{(i)} \cdot w^{(i)}}{\sum_{i=1}^N w^{(i)}}, \quad w^{(i)} = \hat{q}_t^G(x_t \mid x_0^{(i)}) \cdot \frac{\hat{q}^G(x_0^{(i)})}{p(x_0^{(i)} \mid x_t)}. \tag{9}$$

The design of the proposal $p(x_0 \mid x_t)$ aims to ensure that the quotient $\hat{q}^G(x_0)/p(x_0 \mid x_t)$ becomes tractable. This can be achieved by first sampling from the original dataset D using some user-defined $\bar{p}(x_0 \mid x_t)$, where $\bar{p}$ is designed to have non-zero probability for all elements in the dataset. Once a sample $x_0 \in D$ is drawn, we sample a group element g from the group G uniformly at random, and apply this group action to the sample. This results in a new sample $x_0^{(i)} = g \circ x_0$. This proposal distribution inherits the same orbit-integral structure as $\hat{q}^G(x_0)$, causing the intractable terms in the ratio $\hat{q}^G(x_0^{(i)})/p(x_0^{(i)} \mid x_t)$ to cancel. Specifically, with δ the Dirac delta function:

$$\frac{\hat{q}^G(x_0^{(i)})}{p(x_0^{(i)} \mid x_t)} = \frac{\hat{q}^G(g \circ x_0)}{p(g \circ x_0 \mid x_t)} = \frac{\hat{q}(x_0) \int_G \delta(g \circ x_0 - g' \circ x_0) \, \mathrm{d}\mu_G(g')}{\bar{p}(x_0 \mid x_t) \int_G \delta(g \circ x_0 - g' \circ x_0) \, \mathrm{d}\mu_G(g')} = \frac{\hat{q}(x_0)}{\bar{p}(x_0 \mid x_t)}. \tag{10}$$

Beside, since $\hat{q}(x_0) = 1/|D|$ for any $x_0 \in D$, this term can be omitted from the importance weight. Thus, the final importance weight simplifies to $w^{(i)} = \hat{q}_t^G(x_t \mid x_0^{(i)})/\bar{p}(x_0 \mid x_t)$. Importantly, all components of the importance weights are tractable: $\hat{q}_t^G(x_t \mid x_0^{(i)})$ is the forward diffusion process; $\hat{q}(x_0)$ corresponds to the empirical data distribution; and $\bar{p}(x_0 \mid x_t)$ is user-defined and tractable. Moreover, SNIS estimators based on these importance weights are always consistent.

An important instance results from setting $\bar{p}(x_0 \mid x_t) = \hat{q}(x_0)$ where the proposal recovers the exact symmetrized distribution: $p(x_0 \mid x_t) = \hat{q}^G(x_0)$ and the importance weight simplifies to $w^{(i)} = \hat{q}_t^G(x_t \mid x_0^{(i)})$.

3.4 Practical Implementation—Orbit Diffusion (`OrbDiff`)

We present *Orbit Diffusion* (`OrbDiff`) as a practical variant of our estimator. Although using $\hat{q}^G(x_0)$ as the proposal is theoretically valid, it is inefficient and cumbersome in practice. For small t, the conditional $\hat{q}_t^G(x_t \mid x_0)$ is sharply concentrated around the x_0 that generated x_t, so uniformly sampled x_0 rarely yield useful gradients. Furthermore, sampling from the full support generally requires drawing points outside the current minibatch, adding non-trivial implementation complexity.

To improve efficiency, we fix x_0 to the example that produced x_t and sample candidates only from its *orbit* $\mathcal{O}_{x_0} = \{g \circ x_0 \mid g \in G\}$. This biases the proposal toward points with high likelihood under $\hat{q}_t^G(x_t \mid x_0)$—for instance, small rotations in $\mathsf{SO}(3)$-equivariant settings or local permutations in discrete symmetry groups. Since such points dominate the conditional distribution at small noise levels, orbit sampling greatly improves sample efficiency by prioritizing candidates with non-trivial importance weights. At high noise levels, contributions from outside the orbit may increase, but their weights are typically small, and expanding the proposal has shown little benefit.

Formally, `OrbDiff` replaces the intractable conditional expectation $\mathbb{E}[x_0 \mid x_t]$ in Equation (6) with the *orbit-weighted target*

$$\phi^*(x_0, x_t, t) = \frac{1}{Z(x_t, x_0)} \int_G (g \circ x_0)\, \hat{q}_t^G(x_t \mid g \circ x_0)\, \mathrm{d}\mu_G(g), \tag{11}$$

where $Z(x_t, x_0) = \int_G \hat{q}_t^G(x_t \mid g \circ x_0)\, \mathrm{d}\mu_G(g)$ is the normalization constant. This yields the `OrbDiff` loss:

$$\mathcal{L}_t^{\texttt{OrbDiff}}(\phi) = \mathbb{E}_{x_0 \sim \hat{q}^G(x_0)} \mathbb{E}_{x_t \sim \hat{q}_t^G(\cdot \mid x_0)} \left[\left\| \phi(x_t, t) - \phi^*(x_0, x_t, t) \right\|^2 \right]. \tag{12}$$

Figure 1 provides an illustration and pseudo-code of the practical implementation. When ϕ is G-equivariant (see Appendix B.2 and Chen et al. (2024)), the loss in Equation (12) and its gradient are unchanged if x_0 is drawn from the empirical distribution $\hat{q}(x_0)$ rather than $\hat{q}^G(x_0)$.

Even though $\phi^*(x_0, x_t, t) \neq \mathbb{E}[x_0 \mid x_t]$, for an equivariant forward process the gradient of Equation (12) matches that of Equation (5), ensuring that `OrbDiff` yields an unbiased gradient estimate. The orbit-weighted target is also equivariant, providing a training signal aligned with the model's inductive bias. We formally prove both properties in Appendix B.4.

Unbiased gradient and equivariance of `OrbDiff`

Theorem 2. *Let G be a locally compact isometry group acting on data space Ω, and suppose the forward kernels $\hat{q}_t^G(x_t \mid x_0)$ are G-invariant: $\hat{q}_t^G(g \circ x_t \mid g \circ x_0) = \hat{q}_t^G(x_t \mid x_0)$ for all $g \in G$. Then:*

1. *The `OrbDiff` target $\phi^*(x_0, x_t, t)$ satisfies: $\phi^*(x_0, h \circ x_t, t) = h \circ \phi^*(x_0, x_t, t)$ for all $h \in G$.*
2. *The gradient of the `OrbDiff` loss (12) equals that of the ideal loss (5), i.e., `OrbDiff` provides an unbiased gradient estimator.*

We also explore non-uniform sampling of group elements g for approximating the conditional expectation. For small noise, we sample near the identity action, expanding the neighborhood as noise increases. These distributions, $\nu_t(g)$, depend on the noise schedule. While not all groups support closed-form expressions for the density of individual group elements, they exist for the translation group (sampled from a Gaussian) or $\mathsf{SO}(3)$ (sampled from the von Mises-Fisher distribution).

Next, we divide by $\nu_t(g)$ to account for the group sampling distribution, resulting in the importance weights $w^{(i)} = \hat{q}_t^G(x_t \mid x_0^{(i)})/\nu_t(g)$. We also ensure that the identity group element is included in the sampled set. This strategy is effective in practice and provides computational advantages.

4 Experimental Results

Our experiments evaluate the generality and robustness of Orbit Diffusion across diverse generative tasks. We begin with a controlled synthetic setup using a standard diffusion model (Section 4.1) and

consider various isometry groups, including reflections, rotations, translations, and graph automorphisms. We then extend our method to Flow Matching (Section 4.2) and to diffusion models with non-standard forward processes (Section 4.3). Finally, we apply Orbit Diffusion to a non-equivariant denoiser, demonstrating its effectiveness without architectural symmetry (Section 4.4).

4.1 Experiments on Synthetic Data

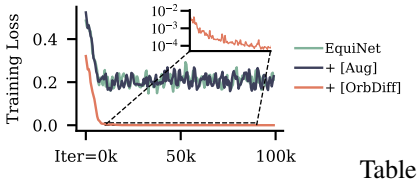


Figure 2: Learning curves.

Variant	RMSD ($\times 10^{-5}$) (↓)	W2 ($\times 10^{-3}$) (↓)
EquiNet	9.05 ± 9.47	1.150 ± 1.581
+ [Aug]	9.50 ± 9.63	0.853 ± 1.156
+ [OrbDiff]	$\mathbf{0.37 \pm 0.09}$	$\mathbf{0.004 \pm 0.001}$

Table 1: Synthetic experiment results: RMSD to the closest target in $\{-1, 1\}$ and W2 distance to ground-truth distribution.

We construct a dataset with a single 1D sample $x_0 = 1$, where the equivariant group is reflection: $g \circ x = \pm x$. We train a denoiser (EquiNet) of the form $D_\theta(x_t, t) = f_\theta(x_t, t) - f_\theta(-x_t, t)$, which is equivariant by design, with f_θ being a simple 3-layer MLP.

We evaluate three training variants of EquiNet: default, trained without Orbit Diffusion or data augmentation; + [Aug], with data augmentation only; and + [OrbDiff], with Orbit Diffusion only. Each model is trained for 100k iterations. After training, we generate 100k samples and compute the root mean square deviation (RMSD) from the closest target in $\{-1, 1\}$, and the Wasserstein-2 (W2) distance to the target distribution. We report mean and standard deviation for each model in Table 1.

Figure 2 shows the training loss curves for all three EquiNet variants. EquiNet and its augmented version exhibit similar fluctuations and magnitudes, indicating that data augmentation does not reduce variance. This is consistent with our theoretical result (Appendix B.2): for an equivariant denoiser, augmented and non-augmented losses are equivalent. In contrast, the OrbDiff variant shows a smoother and lower loss curve, confirming that Rao-Blackwellization reduces gradient variance and stabilizes training. As shown in Table 1, OrbDiff achieves roughly 25× lower RMSD and 200× lower W2 distance, outperforming both EquiNet variants.

4.2 Molecular Conformer Generation

Molecular Conformer Generation. Molecular Conformer Generation aims to generate plausible 3D structures from 2D molecular graphs, crucial for drug discovery and property prediction due to the role of 3D geometry (Liu et al., 2023; Axelrod & Gómez-Bombarelli, 2020).

We evaluate on the GEOM-QM9 dataset (Axelrod & Gomez-Bombarelli, 2022), respecting two key symmetries: invariance under global 3D rotations and equivariance to graph automorphisms, which permute atom indices without altering molecular identity. We compare against strong baselines, including GeoMol (Ganea et al., 2021), Torsional Diffusion (Jing et al., 2022), MCF (Wang et al., 2024b), and ETFlow (Hassan et al., 2024). Our method, Orbit Diffusion, is integrated into ETFlow, a strong equivariant flow matching model that employs a harmonic prior for bonded atom proximity. We finetune ETFLow with OrbDiff using their public checkpoint.

During training, we apply symmetry-aware sampling by uniformly sampling 50 automorphisms and 200 $\mathsf{SO}(3)$ rotations per molecule, including the identity. These are applied to both 2D graphs and 3D conformers. All other settings follow ETFlow; see Appendix C.1 for details.

We benchmark against three versions of ETFlow: the results reported in the original paper, their released checkpoint, and our own reproduced results using the provided code and configuration [3]. Despite extensive effort, we were unable to match their reported performance, so we report all results under the same evaluation protocol.

[3] https://github.com/shenoynikhil/ETFlow

Table 2: Molecular conformer generation performance on GEOM-QM9. * Reported in the original paper. † Obtained using the published checkpoint. ‡ We train the public implementation from scratch.

Models	Recall						Precision					
	Cov@0.1 (↑)		Cov@0.5 (↑)		AMR (↓)		Cov@0.1 (↑)		Cov@0.5 (↑)		AMR (↓)	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GeoMol†	28.4	0.0	91.1	**100.0**	0.224	0.194	20.7	0.0	85.8	**100.0**	0.271	0.243
Torsional Diff.†	37.7	25.0	88.4	**100.0**	0.178	0.147	27.6	12.5	84.5	**100.0**	0.221	0.195
MCF†	81.9	**100.0**	94.9	**100.0**	0.103	0.049	78.6	93.8	**93.9**	**100.0**	**0.113**	0.055
ETFlow*	-	-	96.5	**100.0**	0.073	0.047	-	-	94.1	**100.0**	0.098	0.039
ETFlow†	79.5	**100.0**	93.8	**100.0**	0.096	0.037	74.4	83.3	88.7	**100.0**	0.142	0.066
ETFlow‡	81.4	**100.0**	94.4	**100.0**	0.092	0.039	74.6	85.5	89.1	**100.0**	0.145	0.064
+ [OrbDiff]	**85.4**	**100.0**	**96.3**	**100.0**	**0.074**	**0.027**	**80.2**	**93.9**	91.9	**100.0**	**0.113**	**0.042**

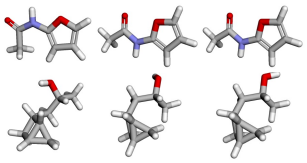


Figure 3: Molecular conformers generated by ETFlow (left), + [OrbDiff] (center), and ground-truth (right).

Table 2 shows our method consistently improves both precision and diversity. OrbDiff achieves the best recall scores, including a 4% improvement in mean Cov@0.1, and the lowest Recall AMR (mean and median). It also maintains competitive precision at 0.1 Å. While MCF performs better at 0.5 Å precision, OrbDiff achieves the lowest AMR overall. Further experimental details and comparisons with more baselines are in Appendix C.1.

4.3 Crystal Structure Prediction (CSP)

CSP involves recovering 3D atomic positions and lattice parameters from chemical composition. Due to periodicity, it suffices to predict the structure within a single unit cell, where coordinates lie in the fractional domain $[0, 1)^{3\times M}$. To handle this, DiffCSP uses a Wrapped Normal diffusion that respects periodic translation symmetry. We integrate Orbit Diffusion into DiffCSP, demonstrating that our approach extends beyond Gaussian diffusion models. We test two variants: OrbDiff_U, which samples uniformly over the translation group, and OrbDiff_WN, a time-dependent Wrapped Normal centered at zero, concentrating around x_0 at low noise and spreading out at high noise. Details of the OrbDiff_WN proposal are in Appendix C.2.3.

We evaluate our method on two CSP benchmarks: Perov-5 (Castelli et al., 2012a,b) and MP-20 (Jain et al., 2013). We use the three strongest baselines from the DiffCSP paper (Jiao et al., 2023): P-cG-SchNet (Gebauer et al., 2022), CDVAE (Xie et al., 2022), and DiffCSP, all with publicly available implementations. We evaluate performance using two standard metrics: Match Rate (the proportion of correctly matched structures in the test set) and RMSD (the average atomic deviation for matched samples, normalized by lattice volume). Full metric definitions and details are in Appendix C.2.4.

We further consider a relevant task, introduced by TGDMat (DAS et al., 2025), where crystal structures are generated conditioned on additional text descriptions of the desired structures. In this task, two types of descriptions are considered: long and short, with the latter being easier to obtain than the former. We follow the same evaluation framework as for Non-text-guided CSP.

Table 3: Text-guided CSP with TGDMat.

Method	Perov-5		MP-20	
	Match (↑)	RMSE (↓)	Match (↑)	RMSE (↓)
TGDMat (S)	59.39	0.066	59.90	0.078
+ [OrbDiff_U]	63.51	0.062	56.50	0.085
+ [OrbDiff_WN]	**65.57**	**0.054**	**61.29**	**0.072**
TGDMat (L)	95.17	0.013	61.91	0.081
+ [OrbDiff_U]	95.88	**0.012**	65.94	**0.069**
+ [OrbDiff_WN]	**95.98**	**0.012**	**66.74**	**0.069**

Table 4: Crystal Structure Prediction (CSP).

Method	Perov-5		MP-20	
	Match (↑)	RMSE (↓)	Match (↑)	RMSE (↓)
P-cG-SchNet	48.22	0.418	15.39	0.376
CDVAE	45.31	0.114	33.90	0.105
DiffCSP	52.02	0.076	51.49	0.063
+ [OrbDiff_U]	52.29	0.078	54.47	0.054
+ [OrbDiff_WN]	**52.39**	**0.069**	**55.70**	**0.053**

From Tables 3 and 4, one can see OrbDiff_WN consistently enhances the performance in all cases, with a notable increase from 59.39% to 65.57% for TGDmat (S) on Perov-5 and from 61.91% to 66.74% for TGDMat (L) on MP-20. At the same time, OrbDiff_U outperforms the baselines in 5 out of 6 cases, showing consistent benefits. Improvements are also observed consistently in RMSE.

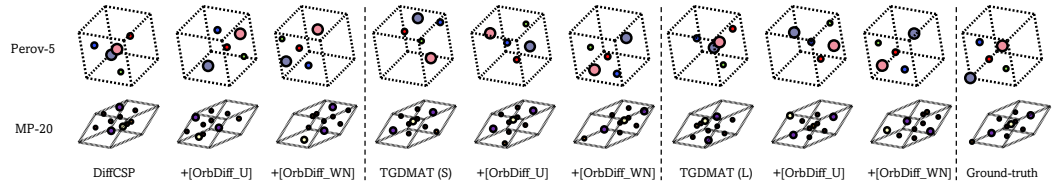


Figure 4: Qualitative comparison of Crystal Structure Predictions by 9 models, including DiffCSP, TGDMat (S) and TGDmat (L) with baselines, `OrbDiff_U`, and `OrbDiff_WN` against ground-truth samples on randomly selected samples from Perov-5 and MP-20 dataset.

4.4 Protein Structure Generation with PROTEINA

Table 5: Protein Structure Generation. Full comparisons are in the appendix. "+ [finetune]" denotes $\mathcal{M}_{FS}^{\text{no-tri}}$ finetuned with the original loss; "+ [`OrbDiff`]" uses `OrbDiff` for finetuning. Full table can be found in Appendix C.3.4

Model	Designability		Diversity		Novelty	
	Fraction (↑)	scRMSD (↓)	Cluster (↑)	TM-score (↓)	PDB (↓)	AFDB (↓)
FoldFlow (OT)	97.2	-	0.37	0.41	0.71	0.75
$\mathcal{M}_{21M}$	**99.0**	**0.72**	0.30	0.39	0.81	0.84
$\mathcal{M}_{FS}^{\text{no-tri}}$	93.8	1.04	**0.62**	**0.36**	**0.69**	**0.76**
+ [finetune]	93.8	1.00	0.54	0.37	0.74	0.83
+ [`OrbDiff`]	**95.6**	**0.93**	0.52	0.37	0.74	0.83

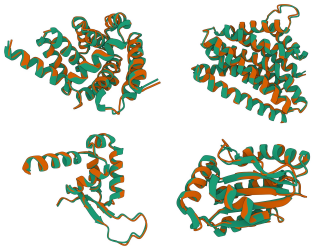


Figure 5: `OrbDiff` (orange) generated structures versus reference structure (green).

We adopt PROTEINA (denoted $\mathcal{M}$) (Geffner et al., 2025), the state-of-the-art model for protein structure generation, as a backbone. Although PROTEINA is a non-equivariant transformer, it performs well through extensive data augmentation. We finetune the 200M-parameter version of PROTEINA—$\mathcal{M}_{FS}^{\text{no-tri}}$—using `OrbDiff`, applying Rao-Blackwellization with a uniform proposal distribution over $\mathrm{SO}(3)$, sampling 10,000 group elements.

For comparison, we also include the best-performing equivariant baseline, FoldFlow (OT) (Bose et al., 2024). We evaluate protein structures using three metrics: Designability (the feasibility of synthesizing the generated structures), Diversity, and Novelty. Designability is the most critical, while Diversity and Novelty are based on the designable samples, ensuring that the generated structures are synthetically plausible, diverse, and novel. Evaluation metric details are in Appendix C.3.3.

Table 5 shows that Orbit Diffusion boosts designability of $\mathcal{M}_{FS}^{\text{no-tri}}$ to 95.6% and lowers scRMSD to 0.93, outperforming naive finetuning while maintaining competitive diversity (Cluster: 0.52) and novelty (PDB: 0.74). The state-of-the-art $\mathcal{M}_{21M}$ (400M parameters) achieves higher designability (99.0%) and scRMSD (0.72) but at the cost of much lower diversity (Cluster: 0.30) and novelty (PDB: 0.81), showing a trade-off between validity and structural variety.

4.5 Benefits of OrbDiff: Efficiency, Stability, and Equivariance

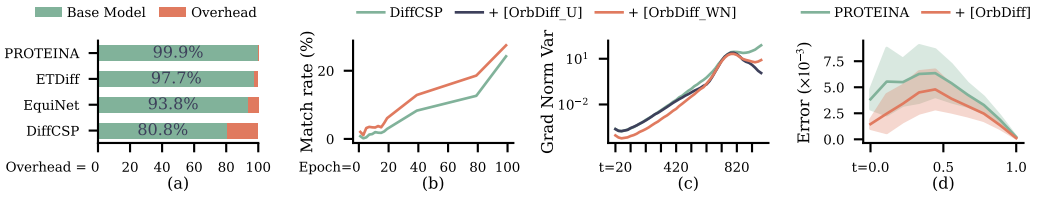


Figure 6: (a) `OrbDiff` introduces minimal computational overhead. (b) Match rate (↑) during DiffCSP training on the Perov-5 dataset: `OrbDiff` accelerates convergence. (c) `OrbDiff` reduces gradient variance across noise levels. (d) `OrbDiff` improves the equivariance of PROTEINA.

Adding `OrbDiff` introduces minimal overhead (Figure 6a)—about 20% for smaller models such as DiffCSP, and only 0.1% for larger ones like PROTEINA (200M parameters). In a 24-hour training run, this corresponds to just 1.4 additional minutes. In practice, the extra memory and computation for sampling and averaging over group elements are negligible compared to the overall training cost. For example, in our PROTEINA experiments—one of the largest protein diffusion models—we use 10,000 SO(3) samples per training example, which adds only $\sim$40 MB per GPU to the total memory usage of $\sim$54 GB.

To better understand how `OrbDiff` accelerates convergence, we compare DiffCSP and DiffCSP + [`OrbDiff`] across training checkpoints. As shown in Figure 6b, integrating `OrbDiff` consistently improves match rates throughout training, especially in the early stages.

We also compare the empirical gradient norm variance of DiffCSP (Perov-5) with `OrbDiff_U` and `OrbDiff_WN` across the full training set and various timesteps. As expected, `OrbDiff_WN` achieves significantly lower variance at small to intermediate noise levels by sampling locally around x_0, while `OrbDiff_U` performs better at high noise levels due to its more global sampling. Both methods substantially reduce variance compared to DiffCSP. Finally, to assess equivariance preservation, we compare PROTEINA + [finetune] and PROTEINA + [`OrbDiff`] using an equivariance test from (Geffner et al., 2025):

$$\text{Error}_t = \mathbb{E}_{x\sim\hat{q}^G(x_0),\ x_t\sim\hat{q}_t^G(x_t|x_0),\ g\sim\text{Unif}(SO(3))}\left[\text{RMSD}(g\circ\phi(x_t,t),\ \phi(g\circ x_t,t))\right]. \tag{13}$$

As shown in Figure 6d, `OrbDiff` substantially reduces equivariance error compared to naive finetuning, indicating improved geometric consistency in the model's denoising.

5 Related Work

Equivariant neural networks have been extensively studied for their ability to encode symmetry priors in domains such as vision (Cohen & Welling, 2016; Worrall et al., 2017), 3D geometry (Thomas et al., 2018; Deng et al., 2021), and molecular modeling (Le et al., 2022; Kozinsky et al., 2023). More recently, these ideas have been incorporated into diffusion models to improve generative performance in structured domains like proteins and molecules (Corso et al., 2024; Hoogeboom et al., 2022; Igashov et al., 2024). By designing the denoising network to be equivariant under a symmetry group, these models better align with the underlying data distribution. A comprehensive discussion of related work is provided in Appendix A.

6 Conclusion

Orbit Diffusion is a framework for training generative models under symmetry constraints by reducing gradient variance through Rao–Blackwellization. The approach unifies equivariant architectures and data augmentation strategies within a single probabilistic formulation, providing a provably lower-variance estimator while maintaining computational efficiency. Theoretically, we show that the proposed loss admits equivariant minimizers and connects to existing score-based and diffusion formulations. Empirically, Orbit Diffusion demonstrates strong and consistent performance across a diverse set of generative tasks, including molecular, crystal structure, and protein structure modeling. By bridging the gap between symmetry-aware modeling and optimization stability, our method improves both the scalability and practical applicability of equivariant generative models in scientific domains.

7 Limitations

To estimate the conditional expectation in our gradient estimator, we employ orbit sampling to reduce variance, which leads to improved performance. While efficient, our sampling scheme inherently constrains the choice of proposal distributions and may limit modeling flexibility, especially for tasks requiring diverse or adaptive noise structures. Furthermore, the number of orbit samples directly affects both computational cost and estimation accuracy, suggesting that dynamically adjusting the sampling strategy during training could yield better efficiency–accuracy trade-offs. Future work could explore adaptive or learned proposal mechanisms to enhance generalization and robustness across broader data regimes.

Acknowledgements

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016; the DARPA ANSR program under award FA8750-23-2-0004; the DARPA CODORD program under award HR00112590089; the NSF grant IIS-1943641; the National University of Singapore Start-up Grant (Award No. SUG-251RES250); and gifts from Adobe Research, Cisco Research, and Amazon. We also acknowledge the support of the Stuttgart Center for Simulation Science (SimTech). VT and MN thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. This work further received partial support from the Diabetes Center Berne and from the Ministry of Science, Research and the Arts Baden-Württemberg through the Artificial Intelligence Software Academy (AISA).

References

Emmanuel Abbe and Enric Boix-Adserà. On the non-universality of deep learning: Quantifying the cost of symmetry. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

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*, 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, 2024b.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *CoRR*, abs/2205.15019, 2022.

Simon Axelrod and Rafael Gómez-Bombarelli. Molecular machine learning with conformer ensembles. *Machine Learning: Science and Technology*, 4, 2020. URL `https://api.semanticscholar.org/CorpusId:229181029`.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 2022.

Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. Se(3)-stochastic flow matching for protein backbone generation. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale? *arXiv preprint arXiv:2410.23179*, 2024.

Ivano E. Castelli, David D. Landis, Kristian S. Thygesen, Søren Dahl, Ib Chorkendorff, Thomas F. Jaramillo, and Karsten W. Jacobsen. New cubic perovskites for one- and two-photon water splitting using the computational materials repository. *Energy & Environmental Science*, 2012a.

Ivano E Castelli, Thomas Olsen, Soumendu Datta, David D Landis, Søren Dahl, Kristian S Thygesen, and Karsten W Jacobsen. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy & Environmental Science*, 5(2):5814–5819, 2012b.

Ziyu Chen, Markos A. Katsoulakis, and Benjamin J. Zhang. Equivariant score-based generative models provably learn distributions with symmetries efficiently. *arXiv preprint arXiv:2410.01244*, 2024.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Gabriele Corso, Arthur Deng, Nicholas Polizzi, Regina Barzilay, and Tommi Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization. In *International Conference on Learning Representations (ICLR)*, 2024.

KISHALAY DAS, Subhojyoti Khastagir, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Periodic materials generation using text-guided joint diffusion model. In *The Thirteenth International Conference on Learning Representations*, 2025.

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, Patrick JY Leung, Thomas F Huddy, Samuel Pellock, Daniel Tischer, Felix Chan, Brian Koepnick, Huy Nguyen, Andrew Kang, Balamurugan Sankaran, Abhishek K Bera, Neil P King, and David Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 2022.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J. Guibas. Vector neurons: A general framework for so(3)-equivariant networks. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Fabian Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Octavian Ganea, Lagnajit Pattanaik, Connor W. Coley, Regina Barzilay, Klavs F. Jensen, William H. Green Jr., and Tommi S. Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021.

Benoit Gaujac, J'er'emie Dona, Liviu Copoiu, Timothy Atkinson, Thomas Pierrot, and Thomas D. Barrett. Learning the language of protein structure. *ArXiv*, abs/2405.15840, 2024. URL `https://api.semanticscholar.org/CorpusId:270063152`.

Niklas W. A. Gebauer, Michael Gastegger, Stefaan S. P. Hessmann, Klaus-Robert Müller, and Kristof T. Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature Communications*, 13(1):973, 2022. doi: 10.1038/s41467-022-28526-y. URL `https://doi.org/10.1038/s41467-022-28526-y`.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Mathis Gerdes, Pim de Haan, Corrado Rainone, Roberto Bondesan, and Miranda C. N. Cheng. Learning lattice quantum field theories with equivariant continuous flows. *SciPost Physics*, 15 (6):238, 2023. doi: 10.21468/SciPostPhys.15.6.238. URL `https://scipost.org/10.21468/SciPostPhys.15.6.238`.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stärk, Stephan Thaler, and Dominique Beaini. Et-flow: Equivariant flow-matching for molecular conformer generation. In *Advances in Neural Information Processing Systems*, 2024.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 373(6557), 2025.

Lingshen He, Yuxuan Chen, Zhengyang Shen, Yiming Dong, Yisen Wang, and Zhouchen Lin. Efficient equivariant network. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Jan Hermann, Zeno Schätzle, and Frank Noé. Deep neural network solution of the electronic schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020.

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.

Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael M. Bronstein, Alexander Tong, and Avishek Joey Bose. Sequence-augmented se(3)-flow matching for conditional protein backbone generation. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, pp. 1–11, 2024.

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, Shan Tie, Vincent Xue, Sarah C. Cowles, Alan Leung, João V. Rodrigues, Claudio L. Morales-Perez, Alex M. Ayoub, Robin Green, Katherine Puentes, Frank Oplinger, Nishant V. Panwar, Fritz Obermeyer, Adam R. Root, Andrew L. Beam, Frank J. Poelwijk, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. *Nature*, 2023.

A Jain, SP Ong, G Hautier, W Chen, WD Richards, S Dacek, S Cholia, D Gunter, D Skinner, G Ceder, et al. The materials project: a materials genome approach to accelerating materials innovation, apl mater. 1 (2013) 011002, NA.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DNdN26m2Jk.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pp. 24240–24253, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/994545b2308bbbbc97e3e687ea9e464f-Abstract-Conference.html.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. In *Machine Learning for Drug Discovery Workshop at the International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/pdf?id=BgbRVzfQqFp.

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.

Nayoung Kim, Seongsu Kim, Minsu Kim, Jinkyoo Park, and Sungsoo Ahn. Mofflow: Flow matching for structure prediction of metal-organic frameworks. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://openreview.net/forum?id=dNT3abOsLo.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36, 2024.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: Exact likelihood generative learning for symmetric densities. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5361–5370. PMLR, 2020. URL http://proceedings.mlr.press/v119/kohler20a.html.

Boris Kozinsky, Albert Musaelian, Anders Johansson, and Simon L. Batzner. Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size. In Dorian Arnold, Rosa M. Badia, and Kathryn M. Mohror (eds.), *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2023, Denver, CO, USA, November 12-17, 2023*, pp. 2:1–2:12. ACM, 2023. doi: 10.1145/3581784.3627041. URL https://doi.org/10.1145/3581784.3627041.

Tuan Le, Frank Noé, and Djork-Arné Clevert. Representation learning on biomolecular structures using equivariant graph attention. In Bastian Rieck and Razvan Pascanu (eds.), *Learning on Graphs Conference, LoG 2022, 9-12 December 2022, Virtual Event*, volume 198 of *Proceedings of Machine Learning Research*, pp. 30. PMLR, 2022. URL https://proceedings.mlr.press/v198/le22a.html.

Tuan Le, Julian Cremer, Frank Noé, Djork-Arné Clevert, and Kristof Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.

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. URL https://arxiv.org/abs/2405.15489.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023.

Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with SE(3)-invariant denoising distance matching. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. URL https://openreview.net/forum?id=CjTHVo1dvR.

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.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Artem R Oganov and Colin W Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics*, 124(24), 2006.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

Xuecheng Shao, Jian Lv, Peng Liu, Sen Shao, P. Gao, Hanyu Liu, Yanchao Wang, and Yanming Ma. A symmetry-orientated divide-and-conquer method for crystal structure prediction. *The Journal of chemical physics*, 156 1:014105, 2021. URL https://api.semanticscholar.org/CorpusId:238198323.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

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 (ICLR)*, 2021b.

Nathaniel Thomas, Tess E. Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. In *Advances in Neural Information Processing Systems (NeurIPS) 31*, 2018.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Vinh Tong, Trung-Dung Hoang, Anji Liu, Guy Van den Broeck, and Mathias Niepert. Learning to discretize denoising diffusion odes. In *Proceedings of the 13th International Conference on Learning Representations*, 2025.

Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Chentong Wang, Yannan Qu, Zhangzhi Peng, Yukai Wang, Hongli Zhu, Dachuan Chen, and Longxing Cao. Proteus: exploring protein structure generation for enhanced designability and efficiency. *bioRxiv*, pp. 2024–02, 2024a.

Yuyang Wang, Ahmed A. A. Elhag, Navdeep Jaitly, Joshua M. Susskind, and Miguel Ángel Bautista. Swallowing the bitter pill: Simplified scalable conformer generation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024b.

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, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Lai Wei, Sadman Sadeed Omee, Rongzhi Dong, Nihang Fu, Yuqi Song, E. Siriwardane, Meiling Xu, Chris Wolverton, and Jianjun Hu. Cspbench: a benchmark and critical evaluation of crystal structure prediction. In *NA*, 2024. URL `https://api.semanticscholar.org/CorpusId:270869939`.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 7168–7177. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.758. URL `https://doi.org/10.1109/CVPR.2017.758`.

Kevin E. Wu, Kevin Kaichuang Yang, Rianne van den Berg, James Zou, Alex X. Lu, and Ava P. Amini. Protein structure generation via folding diffusion. *Nature Communications*, 15, 2022. URL `https://api.semanticscholar.org/CorpusId:252668551`.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022a. URL `https://openreview.net/forum?id=PzcvxEMzvQC`.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022b. URL `https://openreview.net/forum?id=PzcvxEMzvQC`.

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

Jason Yim, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noé. Fast protein backbone generation with se(3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023a.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Viktor Zaverkin, Francesco Alesiani, Takashi Maruyama, Federico Errica, Henrik Christiansen, Makoto Takamoto, Nicolas Weber, and Mathias Niepert. Higher-rank irreducible cartesian tensors for equivariant message passing. In *Conference on Neural Information Processing Systems*, 2024.

Mengchun Zhang, Maryam Qamar, Taegoo Kang, Yuna Jung, Chenshuang Zhang, Sung-Ho Bae, and Chaoning Zhang. A survey on graph diffusion models: Generative AI in science for molecule, protein and material. *CoRR*, abs/2304.01565, 2023. doi: 10.48550/ARXIV.2304.01565. URL `https://doi.org/10.48550/arXiv.2304.01565`.

RAO-BLACKWELL GRADIENT ESTIMATORS FOR EQUIVARIANT DENOISING DIFFUSION

ADDITIONAL MATERIAL

A Related Work

Equivariant Neural Networks. Equivariant neural networks have attracted significant attention for tasks involving structured data, such as computer vision (Cohen & Welling, 2016; Worrall et al., 2017), 3D modeling (Thomas et al., 2018; Fuchs et al., 2020; Deng et al., 2021), quantum mechanics and quantum field theory (Gerdes et al., 2023; Hermann et al., 2020), biomolecular design (Le et al., 2022; Kozinsky et al., 2023; Geffner et al., 2025). These networks exploit group symmetries to ensure consistent outputs under transformations such as rotations, translations, and permutations, which are commonly encountered in many scientific domains. By incorporating symmetric inductive biases, equivariant networks enhance model generalization and reduce data requirements by naturally encoding symmetry constraints. However, challenges remain, such as high computational complexity (He et al., 2021) and difficulties in effectively learning with stochastic gradient descent (SGD) (Abbe & Boix-Adserà, 2022).

Equivariance and Diffusion Models. Diffusion models have become a dominant class of generative models, known for their effectiveness in modeling complex data distributions (Song et al., 2021a; Ho et al., 2020; Karras et al., 2022). They work by gradually adding noise to data in a forward process and learning to reverse this corruption using a denoising network ϕ_θ. Incorporating symmetry, particularly equivariance, into these models has shown significant benefits in domains where data lies on geometric manifolds with known symmetries—such as structural biology (Corso et al., 2024; Yim et al., 2023b; Schneuing et al., 2024; Igashov et al., 2024), molecular modeling (Hoogeboom et al., 2022; Guan et al., 2023; Le et al., 2024), and material design (Xie et al., 2022; Gebauer et al., 2022; Jiao et al., 2023). In such settings, where symmetries are typically governed by the Euclidean group or its subgroups, enforcing equivariance in the generative process helps ensure physical plausibility and improves generalization. A common approach is to design the denoiser ϕ_θ to be equivariant under a symmetry group G, which encourages the learned distribution to converge toward the correct invariant target (Hoogeboom et al., 2022; Xu et al., 2022a; Bose et al., 2024; Chen et al., 2024).

Nonetheless, recent advances have shown that models without explicit equivariant constraints can still achieve strong empirical performance, thanks to greater flexibility in architectural design and effective use of data augmentation. These approaches may implicitly capture symmetry through training strategies rather than architectural bias, as demonstrated by state-of-the-art models such as PROTEINA (Geffner et al., 2025) and AlphaFold 3 (Abramson et al., 2024a).

B Theoretical Proofs

B.1 Proof of Theorem 1

Variance Reduction of Rao-Blackwell estimator.

Theorem 1. *Let $\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)]$ and $\widehat{\nabla}_\phi^{(RB)}[\mathcal{L}_t^G(\phi)]$ denote the gradient estimators of Equation* (4) *and Equation* (5)*, respectively. Suppose we can compute $\mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]$. Then*

$$\mathrm{Var}\left(\widehat{\nabla}_\phi^{(RB)}[\mathcal{L}_t^G(\phi)]\right) \leq \mathrm{Var}\left(\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)]\right).$$

Moreover, the inequality is strict *unless $\hat{q}_t^G(x_0 \mid x_t)$ is a Dirac delta, which is rarely the case in generative modeling where x_0 is typically stochastic given x_t.*

Proof. By definition of the Rao-Blackwellized estimator, we have

$$\widehat{\nabla}_\phi^{(RB)}[\mathcal{L}_t^G(\phi)] = \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}\left[\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)] \mid x_t\right].$$

Thus, the Rao-Blackwell theorem implies that conditioning reduces variance:

$$\mathrm{Var}\left(\mathbb{E}\left[\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)] \mid x_t\right]\right) \leq \mathrm{Var}\left(\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)]\right),$$

with equality if and only if $\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)]$ is almost surely a function of x_t (i.e., deterministic given x_t).

To verify this explicitly, consider the variance of the Rao-Blackwellized estimator:

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\nabla}_\phi^{(RB)}[\mathcal{L}_t^G(\phi)]\right) &= \mathbb{E}_{x_t}\left[\left(\mathbb{E}\left[\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)] \mid x_t\right] - \nabla_\phi[\mathcal{L}_t^G(\phi)]\right)^2\right] \\
&= \mathbb{E}_{x_t}\left[\left(\mathbb{E}\left[\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)] - \nabla_\phi[\mathcal{L}_t^G(\phi)] \mid x_t\right]\right)^2\right] \\
&\leq \mathbb{E}_{x_t}\left[\mathbb{E}\left[\left(\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)] - \nabla_\phi[\mathcal{L}_t^G(\phi)]\right)^2 \mid x_t\right]\right] \\
&= \mathrm{Var}\left(\widehat{\nabla}_\phi[\mathcal{L}_t^G(\phi)]\right),
\end{aligned}$$

where the inequality follows from Jensen's inequality (applied to the convex function $f(z) = z^2$).

Therefore, the Rao-Blackwellized estimator has variance less than or equal to the original estimator, and strictly less unless the conditional variance given x_t is zero. □

B.2 Equivalence of Symmetrized and Equivariant Diffusion Losses

Proposition 1. *Let $\mathcal{L}_t^{(1)}(\phi)$ denote the* Symmetrized Diffusion Loss*, defined as*

$$\mathcal{L}_t^{(1)}(\phi) = \mathbb{E}_{x_0' \sim \hat{q}^G(x_0)} \mathbb{E}_{x_t \sim \hat{q}_t^G(x_t|x_0')}\left[\left\|\phi(x_t, t) - \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2\right],$$

and let $\mathcal{L}_t^{(2)}(\phi)$ denote the corresponding loss defined on the original (non-symmetrized) data distribution $\hat{q}(x_0)$:

$$\mathcal{L}_t^{(2)}(\phi) = \mathbb{E}_{x_0' \sim \hat{q}(x_0)} \mathbb{E}_{x_t \sim \hat{q}_t(x_t|x_0')}\left[\left\|\phi(x_t, t) - \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2\right].$$

Suppose $\phi(x_t, t)$ is a G-equivariant function, i.e.,

$$\phi(g \circ x_t, t) = g \circ \phi(x_t, t) \quad \forall\, g \in G,\ x_t \in \Omega.$$

Then $\mathcal{L}_t^{(1)}(\phi)$ and $\mathcal{L}_t^{(2)}(\phi)$ are equivalent in the sense that they share the same minimizer and gradient with respect to ϕ.

To prove this, we need the following lemma:

Symmetrized Forward Diffusion Distributions

Lemma 1. *Let $\hat{q}(x_0)$ be an empirical distribution and $\hat{q}^G(x_0)$ its symmetrized counterpart under a symmetry group G, defined by $\hat{q}^G(x_0) = S_G[\hat{q}](x_0)$. Suppose a forward diffusion process acting on $\hat{q}(x_0)$ yields time-dependent marginal distributions $\hat{q}_t(x_t)$. Let a similar process act on $\hat{q}^G(x_0)$ to generate $\hat{q}_t^G(x_t)$. Then, for all $t \geq 0$, the following holds:*

$$\hat{q}_t^G(x_t) = S_G[\hat{q}_t](x_t).$$

The proof can be found in Appendix B.3.

We also need the following lemma:

Lemma 2. *Let G be a group of isometries acting on Ω, and suppose the distribution $\hat{q}_t^G(x_t \mid x_0)$ is equivariant under the action of G. Then for any $g \in G$, the following identity holds:*

$$\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0|g \circ x_t)}[x_0] = g \circ \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0|x_t)}[x_0].$$

Proof. We start by expanding the conditional expectation:

$$\begin{aligned}\mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|g\circ x_t)}[x_0] &= \int_\Omega x_0\,\hat{q}_t^G(x_0 \mid g\circ x_t)\,\mathrm{d}x_0 \\ &= \int_\Omega x_0\,\hat{q}_t^G(g\circ x_t \mid x_0)\frac{\hat{q}^G(x_0)}{\hat{q}_t^G(g\circ x_t)}\,\mathrm{d}x_0.\end{aligned}$$

Now, apply the change of variable $x_0 = g\circ\bar{x}_0$. Since g is an isometry, the Lebesgue measure is invariant, i.e., $\mathrm{d}x_0 = \mathrm{d}\bar{x}_0$. Therefore:

$$\begin{aligned}&= \int_\Omega (g\circ\bar{x}_0)\,\hat{q}_t^G(g\circ x_t \mid g\circ\bar{x}_0)\frac{\hat{q}^G(g\circ\bar{x}_0)}{\hat{q}_t^G(g\circ x_t)}\,\mathrm{d}\bar{x}_0 \\ &= g\circ\int_\Omega \bar{x}_0\,\hat{q}_t^G(x_t \mid \bar{x}_0)\frac{\hat{q}^G(\bar{x}_0)}{\hat{q}_t^G(x_t)}\,\mathrm{d}\bar{x}_0 \\ &= g\circ\mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0],\end{aligned}$$

where the second equality follows from the equivariance of $\hat{q}_t^G$ and the invariance of $\hat{q}^G$ under the group action. This concludes the proof. □

Now the prove the Proposition 1.

Proof. We begin by simplifying $\mathcal{L}_t^{(1)}$:

$$\begin{aligned}\mathcal{L}_t^{(1)}(\phi) &= \mathbb{E}_{x_t\sim\hat{q}_t^G(x_t)}\mathbb{E}_{x_0'\sim\hat{q}^G(x_0|x_t)}\left[\left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2\right] \\ &= \mathbb{E}_{x_t\sim\hat{q}_t^G(x_t)}\left[\left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2\right].\end{aligned}$$

Similarly,

$$\mathcal{L}_t^{(2)}(\phi) = \mathbb{E}_{x_t\sim\hat{q}_t(x_t)}\left[\left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2\right].$$

We now rewrite $\mathcal{L}_t^{(1)}$ in integral form:

$$\mathcal{L}_t^{(1)}(\phi) = \int_\Omega \left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2 \hat{q}_t^G(x_t)\,\mathrm{d}x_t.$$

Using the decomposition of the measure over the group action, we change variables:

$$\int_\Omega f(x_t)\,\mathrm{d}x_t = \int_{\Omega/G}\int_G f(g\circ x_t)\,\mathrm{d}\mu_G(g)\,\mathrm{d}x_t.$$

Thus,

$$\begin{aligned}\mathcal{L}_t^{(1)}(\phi) &= \int_{\Omega/G}\int_G \left\|\phi(g\circ x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|g\circ x_t)}[x_0]\right\|^2 \hat{q}_t^G(g\circ x_t)\,\mathrm{d}\mu_G(g)\,\mathrm{d}x_t \\ &= \int_{\Omega/G}\int_G \left\|g\circ\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|g\circ x_t)}[x_0]\right\|^2 \hat{q}_t^G(x_t)\,\mathrm{d}\mu_G(g)\,\mathrm{d}x_t,\end{aligned}$$

where we used the G-equivariance of ϕ and the invariance of $\hat{q}_t^G$.

Since G acts isometrically, we apply:

$$\left\|g\circ\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|g\circ x_t)}[x_0]\right\|^2 = \left\|\phi(x_t,t) - g^{-1}\circ\mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|g\circ x_t)}[x_0]\right\|^2.$$

Next, applying Lemma 2, we have:

$$\left\|g\circ\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|g\circ x_t)}[x_0]\right\|^2 = \left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2$$

We conclude:

$$\begin{aligned}\mathcal{L}_t^{(1)}(\phi) &= \int_{\Omega/G}\int_G \left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2 \hat{q}_t^G(x_t)\,\mathrm{d}\mu_G(g)\,\mathrm{d}x_t \\ &= \int_{\Omega/G} \left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2 \hat{q}_t^G(x_t)\,\mathrm{d}x_t.\end{aligned}$$

Next, we similarly transform $\mathcal{L}_t^{(2)}(\phi)$:

$$\begin{aligned}\mathcal{L}_t^{(2)}(\phi) &= \int_{\Omega/G}\int_G \left\|\phi(g\circ x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|g\circ x_t)}[x_0]\right\|^2 \hat{q}_t(g\circ x_t)\,\mathrm{d}\mu_G(g)\,\mathrm{d}x_t \\ &= \int_{\Omega/G} \left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2 \left(\int_G \hat{q}_t(g\circ x_t)\,\mathrm{d}\mu_G(g)\right)\mathrm{d}x_t.\end{aligned}$$

Using the theoretical result from Appendix B.3, we have:

$$\hat{q}_t^G(x_t) = S_G[\hat{q}_t](x_t) = \int_G \hat{q}_t(g\circ x_t)\,\mathrm{d}\mu_G(g),$$

we conclude:

$$\mathcal{L}_t^{(2)}(\phi) = \int_{\Omega/G} \left\|\phi(x_t,t) - \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]\right\|^2 \hat{q}_t^G(x_t)\,\mathrm{d}x_t = \mathcal{L}_t^{(1)}(\phi).$$

Hence, the two loss functions are equivalent in the sense that they yield the same gradients and minimizers with respect to ϕ. □

B.3 Symmetrized Forward Diffusion Distribution.

Below is the formal lemma and the proof for the symmetrized forward diffusion distribution.

Symmetrized Forward Diffusion Distributions

Lemma 1. *Let $\hat{q}(x_0)$ be an empirical distribution and $\hat{q}^G(x_0)$ its symmetrized counterpart under a symmetry group G, defined by $\hat{q}^G(x_0) = S_G[\hat{q}](x_0)$. Suppose a forward diffusion process acting on $\hat{q}(x_0)$ yields time-dependent marginal distributions $\hat{q}_t(x_t)$. Let a similar process act on $\hat{q}^G(x_0)$ to generate $\hat{q}_t^G(x_t)$. Then, for all $t \geq 0$, the following holds:*

$$\hat{q}_t^G(x_t) = S_G[\hat{q}_t](x_t).$$

Proof. The marginal distribution at time step t of a diffusion process is defined as

$$\hat{q}_t(x_t) = \int_\Omega q_t(x_t \mid x_0)\hat{q}(x_0)\,\mathrm{d}x_0,$$

where $q_t(x_t \mid x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 I)$ is the Gaussian diffusion kernel, which is equivariant under isometry group transformations. Specifically, for any $g \in G$, we have

$$q_t(x_t \mid x_0) = q_t(g\circ x_t \mid g\circ x_0).$$

The symmetrized marginal distribution at time t is defined as

$$\begin{aligned}S_G[\hat{q}_t](x_t) &= \int_G \hat{q}_t(g\circ x_t)\mathrm{d}\mu_G(g) \\ &= \int_G \left[\int_\Omega q_t(g\circ x_t \mid x_0)\hat{q}(x_0)\,\mathrm{d}x_0\right]\mathrm{d}\mu_G(g) \\ &= \int_G\int_\Omega q_t(g\circ x_t \mid x_0)\hat{q}(x_0)\,\mathrm{d}x_0\mathrm{d}\mu_G(g).\end{aligned}$$

Next, we compute the marginal distribution at time t when the forward process is applied to the symmetrized data distribution:

$$\begin{aligned}\hat{q}_t^G(x_t) &= \int_\Omega q_t(x_t \mid x_0) S_G[\hat{q}](x_0)\,\mathrm{d}x_0 \\ &= \int_\Omega q_t(x_t \mid x_0) \int_G \hat{q}(g \circ x_0)\mathrm{d}\mu_G(g)\mathrm{d}x_0 \\ &= \int_\Omega \int_G q_t(x_t \mid x_0)\hat{q}(g \circ x_0)\mathrm{d}\mu_G(g)\mathrm{d}x_0 \\ &= \int_G \int_\Omega q_t(x_t \mid x_0)\hat{q}(g \circ x_0)\mathrm{d}x_0\mathrm{d}\mu_G(g).\end{aligned}$$

Applying a change of variable $x_0 \mapsto g^{-1} \circ x_0$, we get

$$\begin{aligned}\hat{q}_t^G(x_t) &= \int_G \int_\Omega q_t(x_t \mid g^{-1} \circ x_0)\hat{q}(g \circ [g^{-1} \circ x_0])\mathrm{d}(g^{-1} \circ x_0)\mathrm{d}\mu_G(g) \\ &= \int_G \int_\Omega q_t(x_t \mid g^{-1} \circ x_0)\hat{q}(x_0)\mathrm{d}x_0\mathrm{d}\mu_G(g) \quad \text{(since the Jacobian of } g \text{ is 1)} \\ &= \int_G \int_\Omega q_t(g \circ x_t \mid x_0)\hat{q}(x_0)\mathrm{d}x_0\mathrm{d}\mu_G(g) \quad \text{(by kernel equivariance).}\end{aligned}$$

Thus, we have

$$\hat{q}_t^G(x_t) = S_G[\hat{q}_t](x_t).$$

This completes the proof. □

B.4 Proof of Theorem 2

Unbiased gradient and equivariance of `OrbDiff`

Theorem 2. *Let G be a locally compact isometry group acting on data space Ω, and suppose the forward kernels $\hat{q}_t^G(x_t \mid x_0)$ are G-invariant: $\hat{q}_t^G(g \circ x_t \mid g \circ x_0) = \hat{q}_t^G(x_t \mid x_0)$ for all $g \in G$. Then:*

1. *The `OrbDiff` target $\phi^*(x_0, x_t, t)$ satisfies: $\phi^*(x_0, h \circ x_t, t) = h \circ \phi^*(x_0, x_t, t)$ for all $h \in G$.*
2. *The gradient of the `OrbDiff` loss* (12) *equals that of the ideal loss* (5)*, i.e., `OrbDiff` provides an unbiased gradient estimator.*

We first prove that the `OrbDiff` target is equivariant:

Proof. We compute $\phi^*(x_0, h \circ x_t, t)$ using the definition:

$$\phi^*(x_0, h \circ x_t, t) = \frac{1}{Z(h \circ x_t, x_0)} \int_G (g \circ x_0)\, \hat{q}_t^G(h \circ x_t \mid g \circ x_0)\, \mathrm{d}\mu_G(g).$$

By the equivariance of $\hat{q}_t^G$, we have:

$$\hat{q}_t^G(h \circ x_t \mid g \circ x_0) = \hat{q}_t^G(x_t \mid h^{-1} \circ g \circ x_0).$$

Letting $g' = h^{-1} \circ g$, so $g = h \circ g'$, and using the left-invariance of the Haar measure μ_G, we get:

$$\phi^*(x_0, h \circ x_t, t) = \frac{1}{Z(h \circ x_t, x_0)} \int_G (h \circ g' \circ x_0)\, \hat{q}_t^G(x_t \mid g' \circ x_0)\, \mathrm{d}\mu_G(g').$$

Factoring out h from the integrand gives:

$$= h \circ \left[\frac{1}{Z(h \circ x_t, x_0)} \int_G (g' \circ x_0)\, \hat{q}_t^G(x_t \mid g' \circ x_0)\, \mathrm{d}\mu_G(g') \right].$$

It remains to show that $Z(h \circ x_t, x_0) = Z(x_t, x_0)$, where:

$$Z(x_t, x_0) := \int_G \hat{q}_t^G(x_t \mid g \circ x_0) \, \mathrm{d}\mu_G(g).$$

Using the same substitution:

$$Z(h \circ x_t, x_0) = \int_G \hat{q}_t^G(h \circ x_t \mid g \circ x_0) \, \mathrm{d}\mu_G(g) = \int_G \hat{q}_t^G(x_t \mid h^{-1} \circ g \circ x_0) \, \mathrm{d}\mu_G(g) = \int_G \hat{q}_t^G(x_t \mid g' \circ x_0) \, \mathrm{d}\mu_G(g') = Z(x_t).$$

Therefore,

$$\phi^*(x_0, h \circ x_t, t) = h \circ \phi^*(x_0, x_t, t).$$

□

Next, we prove that `OrbDiff` yields an unbiased gradient estimator:

Proof. Our goal is to estimate the following gradient:

$$\nabla_\phi \mathcal{L}_t^G(\phi) = 2\mathbb{E}_{x_t \sim \hat{q}_t^G(x_t)} \left[\phi(x_t, t) - \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)}[x_0] \right]. \tag{14}$$

Our proposed Rao-Blackwell loss function has the same gradient as $\mathcal{L}_t^G(\phi)$, but with reduced variance due to the use of the conditional expectation $\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)}[x_0]$:

$$\mathcal{L}_t^{\mathrm{RB}}(\phi) = \mathbb{E}_{x_0' \sim \hat{q}^G(x_0')} \mathbb{E}_{x_t \sim \hat{q}_t^G(x_t \mid x_0')} \left[\left\| \phi(x_t, t) - \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)}[x_0] \right\|^2 \right]. \tag{15}$$

However, computing $\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)}[x_0]$ can be computationally expensive. To address this, we introduce OrbDiff, which uses a biased proposal distribution to approximate this expectation using only samples from the orbit of the x_0 that generated x_t. This yields the alternative target:

$$\phi^*(x_0, x_t, t) = \frac{1}{Z(x_0, x_t)} \int_G (g \circ x_0) \, \hat{q}_t^G(g \circ x_0 \mid x_t) \, \mathrm{d}\mu_G(g), \tag{16}$$

where the normalization constant is

$$Z(x_0, x_t) = \int_G \hat{q}_t^G(g \circ x_0 \mid x_t) \, \mathrm{d}\mu_G(g). \tag{17}$$

This matches Equation (11), which can be verified via straightforward transformations.

Although ϕ^* differs from $\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)}[x_0]$, we show that replacing the Rao-Blackwell target in Eq. (15) with ϕ^* yields a loss whose gradient still matches the original gradient. This relies on the assumption that the forward conditional distribution is equivariant under the group action, i.e.,

$$\hat{q}_t^G(g \circ x_t \mid g \circ x_0) = \hat{q}_t^G(x_t \mid x_0), \quad \forall g \in G, \tag{18}$$

which is a natural condition satisfied in many generative models such as diffusion models or flow matching with isotropic Gaussian priors.

Under this assumption, the OrbDiff loss is given by:

$$\mathcal{L}_t^{\mathrm{OrbDiff}}(\phi) = \mathbb{E}_{x_0 \sim \hat{q}^G(x_0)} \mathbb{E}_{x_t \sim \hat{q}_t^G(x_t \mid x_0)} \left[\|\phi(x_t, t) - \phi^*(x_0, x_t, t)\|^2 \right] \tag{19}$$

$$= \mathbb{E}_{x_t \sim \hat{q}_t^G(x_t)} \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)} \left[\|\phi(x_t, t) - \phi^*(x_0, x_t, t)\|^2 \right]. \tag{20}$$

Taking the gradient with respect to ϕ, we obtain:

$$\nabla_\phi \mathcal{L}_t^{\mathrm{OrbDiff}}(\phi) = 2\mathbb{E}_{x_t \sim \hat{q}_t^G(x_t)} \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)} \left[\phi(x_t, t) - \phi^*(x_0, x_t, t) \right] \tag{21}$$

$$= 2\mathbb{E}_{x_t \sim \hat{q}_t^G(x_t)} \left[\phi(x_t, t) - \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)}[\phi^*(x_0, x_t, t)] \right]. \tag{22}$$

Thus, to ensure that OrbDiff yields the correct gradient, it suffices to show:

$$\mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)}[\phi^*(x_0, x_t, t)] = \mathbb{E}_{x_0 \sim \hat{q}_t^G(x_0 \mid x_t)}[x_0]. \tag{23}$$

We compute:

$$\mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[\phi^*(x_0,x_t,t)] = \int_\Omega\int_G \phi^*(g'\circ x_0,x_t,t)\,\hat{q}_t^G(g'\circ x_0\mid x_t)\,\mathrm{d}\mu_G(g')\,\mathrm{d}x_0. \tag{24}$$

Substituting the expression for ϕ^* and applying the change of variables $g \mapsto g\cdot g'$ with left-invariant Haar measure μ_G, we have:

$$= \int_\Omega\int_G \left[\frac{1}{Z(g'\circ x_0,x_t)}\int_G (g\circ[g'\circ x_0])\,\hat{q}_t^G(g\circ[g'\circ x_0]\mid x_t)\,\mathrm{d}\mu_G(g)\right]\hat{q}_t^G(g'\circ x_0\mid x_t)\,\mathrm{d}\mu_G(g')\,\mathrm{d}x_0 \tag{25}$$

$$= \int_\Omega\int_G \left[\frac{1}{Z(x_0,x_t)}\int_G (g\circ x_0)\,\hat{q}_t^G(g\circ x_0\mid x_t)\,\mathrm{d}\mu_G(g)\right]\hat{q}_t^G(g'\circ x_0\mid x_t)\,\mathrm{d}\mu_G(g')\,\mathrm{d}x_0 \tag{26}$$

$$= \int_\Omega \left[\frac{1}{Z(x_0,x_t)}\int_G (g\circ x_0)\,\hat{q}_t^G(g\circ x_0\mid x_t)\,\mathrm{d}\mu_G(g)\right]\left[\int_G \hat{q}_t^G(g'\circ x_0\mid x_t)\,\mathrm{d}\mu_G(g')\right]\mathrm{d}x_0 \tag{27}$$

$$= \int_\Omega \left[\frac{1}{Z(x_0,x_t)}\int_G (g\circ x_0)\,\hat{q}_t^G(g\circ x_0\mid x_t)\,\mathrm{d}\mu_G(g)\right]Z(x_0,x_t)\,\mathrm{d}x_0 \tag{28}$$

$$= \int_\Omega\int_G (g\circ x_0)\,\hat{q}_t^G(g\circ x_0\mid x_t)\,\mathrm{d}\mu_G(g)\,\mathrm{d}x_0 \tag{29}$$

$$= \mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}[x_0]. \tag{30}$$

Therefore, despite ϕ^* not being equal to the conditional expectation at each x_t, the gradient induced by the OrbDiff loss matches the desired gradient. OrbDiff thus provides an unbiased estimate of $\nabla_\phi\mathcal{L}_t^G(\phi)$, while using only samples from the orbit of x_0.

□

B.5 Unconstrained Non-Symmetrized Diffusion Minimizer is not guaranteed G-equivaraint

Lemma 2. *Let*

$$\mathcal{L}_t(\phi) = \mathbb{E}_{(x_0,x_t)\sim\hat{q}(x_0,x_t)}\left[\|\phi(x_t,t)-x_0\|^2\right]$$

be the unconstrained Non-Symmetrized Diffusion loss, where $\hat{q}(x_0,x_t)$ is the empirical joint distribution of clean and noisy data. Then the minimizer $\phi^*(x_t,t)$ of $\mathcal{L}_t$ is the conditional expectation:

$$\phi^*(x_t,t) = \mathbb{E}_{x_0\sim\hat{q}_t(x_0|x_t)}[x_0].$$

*However, this minimizer is not guaranteed to be equivariant under the action of a symmetry group G-*equivariant*.*

Proof. To find the minimizer of the diffusion loss, we first compute the stationary point. The loss is given by:

$$\mathcal{L}_t(\phi) = \mathbb{E}_{(x_0,x_t)\sim\hat{q}(x_0,x_t)}\left[\|\phi(x_t,t)-x_0\|^2\right] = \mathbb{E}_{x_t\sim\hat{q}_t(x_t)}\mathbb{E}_{x_0\sim\hat{q}_t(x_0|x_t)}\left[\|\phi(x_t,t)-x_0\|^2\right].$$

The gradient of the internal expectation is:

$$\begin{aligned}\nabla_\phi\mathbb{E}_{x_0\sim\hat{q}_t(x_0|x_t)}\left[\|\phi(x_t,t)-x_0\|^2\right] &= 2\mathbb{E}_{x_0\sim\hat{q}_t(x_0|x_t)}\left[\phi(x_t,t)-x_0\right]\\ &= 2\left(\phi(x_t,t)-\mathbb{E}_{x_0\sim\hat{q}_t(x_0|x_t)}[x_0]\right).\end{aligned}$$

Setting the gradient to zero gives the minimizer:

$$\phi^*(x_t,t) = \mathbb{E}_{x_0\sim\hat{q}_t(x_0|x_t)}[x_0] = \int_\Omega x_0\hat{q}_t(x_0\mid x_t)\,dx_0.$$

Next, we provide a counterexample to show that $\phi^*(x_t,t)$ is not guaranteed to be equivariant.

Counterexample: Translation in 1D

1. **Data**: Two points $x_0^1 = 0$ and $x_0^2 = 1$, with uniform empirical distribution $\hat{q}(x_0^i) = 0.5$.
2. **Group action**: Translation by $a = 1$, i.e., $g \circ x = x + 1$.
3. **Diffusion kernel**: $q_t(x_t \mid x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2)$.

First, rewrite the minimizer:

$$\begin{aligned}\phi^*(x_t, t) &= \sum_{i=1}^{N} x_0^i \hat{q}_t(x_0^i \mid x_t) \\ &= \sum_{i=1}^{N} x_0^i \hat{q}_t(x_t \mid x_0^i) \frac{\hat{q}(x_0^i)}{\hat{q}_t(x_t)} \\ &= \frac{1}{\hat{q}_t(x_t)} \sum_{i=1}^{N} x_0^i \hat{q}_t(x_t \mid x_0^i) \hat{q}(x_0^i).\end{aligned}$$

Substituting $x_0^1 = 0$ and $x_0^2 = 1$:

$$\begin{aligned}\phi^*(x_t, t) &= \frac{1}{\hat{q}_t(x_t)} \left(0 \cdot \mathcal{N}(x_t; 0, \sigma_t^2) \cdot 0.5 + 1 \cdot \mathcal{N}(x_t; \alpha_t, \sigma_t^2) \cdot 0.5\right), \\ &= \frac{0.5 \cdot \mathcal{N}(x_t; \alpha_t, \sigma_t^2)}{\hat{q}_t(x_t)} = \frac{\mathcal{N}(x_t; \alpha_t, \sigma_t^2)}{\mathcal{N}(x_t; 0, \sigma_t^2) + \mathcal{N}(x_t; \alpha_t, \sigma_t^2)}.\end{aligned}$$

Now, compute $\phi^*(g \circ x_t, t)$ by applying $g \circ x_t = x_t + 1$:

$$\phi^*(g \circ x_t, t) = \frac{\mathcal{N}(x_t + 1; \alpha_t, \sigma_t^2)}{\mathcal{N}(x_t + 1; 0, \sigma_t^2) + \mathcal{N}(x_t + 1; \alpha_t, \sigma_t^2)} < 1.$$

Next, compute $g \circ \phi^*(x_t, t)$:

$$g \circ \phi^*(x_t, t) = \frac{\mathcal{N}(x_t; \alpha_t, \sigma_t^2)}{\mathcal{N}(x_t; 0, \sigma_t^2) + \mathcal{N}(x_t; \alpha_t, \sigma_t^2)} + 1 > 1.$$

Thus,

$$\phi^*(g \circ x_t, t) < g \circ \phi^*(x_t, t),$$

since $\phi^*(g \circ x_t, t) < 1$ and $g \circ \phi^*(x_t, t) > 1$. Consequently,

$$\phi^*(g \circ x_t, t) \neq g \circ \phi^*(x_t, t).$$

This completes the counterexample, showing that $\phi^*(x_t, t)$ is not necessarily equivariant. □

C Experimental Details

C.1 Molecular Conformer Generation (MCG) with Flow Matching

Molecular conformer generation is a fundamental task in computational chemistry and drug discovery, where the goal is to generate plausible 3D structures (conformers) that correspond to a 2D molecular graph. Molecular conformer generation is essential for drug discovery and molecular property prediction, as the 3D structure greatly influences chemical behavior and interactions (Liu et al., 2023; Axelrod & Gómez-Bombarelli, 2020).

To model this task effectively, it is crucial to respect the underlying symmetries of molecular structures. Since conformers are invariant under global rotations and translations, one might consider the full Euclidean group $\mathsf{SE}(3)$. However, in practice, molecular structures are typically zero-centered, effectively removing the need to model translation invariance. As a result, it suffices to consider equivariance under the rotation group $\mathsf{SO}(3)$. In addition, the molecular graph may exhibit symmetry under automorphisms—permutations of atoms that preserve the graph structure—making it important to account for graph isomorphism to avoid redundant representations and ensure physically meaningful predictions.

C.1.1 MCG - Dataset

We evaluate our method on the GEOM-QM9 dataset (Axelrod & Gomez-Bombarelli, 2022), a widely used subset of GEOM containing molecules with an average of 11 atoms. We follow the same train/validation/test split as in (Ganea et al., 2021; Jing et al., 2022), consisting of 106,586 / 13,323 / 1,000 molecules, respectively.

C.1.2 MCG - Baselines

We compare against strong recent baselines with publicly available code, including GEODIFF (Xu et al., 2022b), GEOMOL (Ganea et al., 2021), Torsional Diffusion (Jing et al., 2022), MCF (Wang et al., 2024b), and ETFLOW (Hassan et al., 2024). GEODIFF generates structures using a roto-translationally invariant diffusion process, starting from an invariant initial density and evolving through a Markov kernel that preserves this invariance. GEOMOL predicts 3D structures by modeling torsion angles conditioned on a molecular graph, offering fast inference and good geometric validity. Torsional Diffusion generates conformers using a diffusion process in torsion space. MCF directly models 3D coordinates using a diffusion model without enforcing equivariance, relying instead on model scale to achieve strong performance. ETFLOW, the strongest of these baselines, is an equivariant flow matching model that uses a harmonic prior to encourage spatial proximity of bonded atoms. We integrate Orbit Diffusion into ETFLOW to build on its strong geometric foundation.

C.1.3 MCG - ETFLOW with Orbit Diffusion

While Orbit Diffusion is framed within the context of diffusion models, it naturally extends to flow matching. We introduce this extension through the design of ETFLOW, which employs a harmonic prior and a flexible coupling to the data distribution. Unlike diffusion models, which fix the prior and reverse-time coupling, flow matching allows arbitrary choices for both (Tong et al., 2024), offering greater flexibility in the generative process.

Assume a coupling $q(x_0, x_1)$ between the base distribution q_0 and the data distribution q_1. For each pair (x_0, x_1), define the linear interpolation:

$$I_t(x_0, x_1) = (1-t)x_0 + tx_1, \quad t \in [0, 1].$$

Note: In contrast to diffusion models (where $t = 0$ corresponds to data and $t = 1$ to noise), flow matching treats $x_0 \sim q_0$ as the prior sample and $x_1 \sim q_1$ as the data sample.

ETFLOW defines the conditional distribution:

$$q_t(x_t \mid x_0, x_1) = \mathcal{N}\left(x_t \mid I_t(x_0, x_1), \sigma^2 t(1-t)\right),$$

with small σ, inducing the following velocity field:

$$v_t(x_t) = x_1 - x_0 + \frac{1-2t}{2\sqrt{t(1-t)}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

Given the sampling equation:

$$x_t = (1-t)x_0 + tx_1 + \sigma\sqrt{t(1-t)}\epsilon,$$

we can express ϵ as:

$$\epsilon = \frac{x_t - (1-t)x_0 - tx_1}{\sigma\sqrt{t(1-t)}}.$$

Substituting this into $v_t(x_t)$ yields:

$$\begin{aligned} v_t(x_t) &= x_1 - x_0 + \frac{1-2t}{2\sigma t(1-t)}\left(x_t - (1-t)x_0 - tx_1\right) \\ &= \frac{1-2t}{2\sigma t(1-t)}x_t - \left(1 + \frac{1-2t}{2\sigma t}\right)x_0 + \left(1 - \frac{1-2t}{2\sigma(1-t)}\right)x_1 \\ &= h(t)x_t - g(t)x_0 + f(t)x_1. \end{aligned}$$

The model is trained to match this target velocity using the loss:

$$\mathcal{L}_t(\phi) = \mathbb{E}_{(x_0,x_1)}\, \mathbb{E}_{x_t \sim q_t(\cdot|x_0,x_1)} \left[\|\phi(x_t,t) - v_t(x_t)\|^2 \right],$$

with gradient:

$$\nabla_\phi \mathcal{L}(\phi) = \mathbb{E}_{(x_0,x_1)}\, \mathbb{E}_{x_t \sim q_t(\cdot|x_0,x_1)} \left[2(\phi(x_t,t) - v_t(x_t)) \right].$$

A Rao-Blackwellized gradient can be derived by conditioning on x_t:

$$\begin{aligned}\nabla_\phi \mathcal{L}(\phi) &= \mathbb{E}_{x_t}\left[2\left(\phi(x_t,t) - \mathbb{E}_{(x_0,x_1)|x_t}\left[h(t)x_t - g(t)x_0 + f(t)x_1\right]\right)\right] \\ &= \mathbb{E}_{x_t}\left[2\left(\phi(x_t,t) - h(t)x_t + g(t)\mathbb{E}_{x_0|x_t}[x_0] - f(t)\mathbb{E}_{x_1|x_t}[x_1]\right)\right]\end{aligned}$$

We use a single sample x_0 to estimate $\mathbb{E}[x_0 \mid x_t]$ then use the same estimation technique as in Orbit Diffusion to compute $\mathbb{E}[x_1 \mid x_t]$, enabling efficient Rao-Blackwellized gradient estimation.

C.1.4 MCG - Training Protocol

Instead of training from scratch, we finetune ETFlow using their public checkpoint. During training, we explicitly incorporate both forms of symmetry relevant to molecular data: discrete graph automorphisms and continuous spatial rotations. For each molecule, we uniformly sample 50 elements from its automorphism group using the `pynauty` library, capturing permutation symmetries in the 2D molecular graph structure. Simultaneously, we sample 200 elements from the rotation group $\mathsf{SO}(3)$ to account for the continuous rotational symmetries of its 3D conformation. This symmetry-aware augmentation is applied consistently across the dataset to ensure that the model learns to respect and exploit both types of equivariances. All other training settings, including optimizer configurations and learning rate schedules, follow the defaults of ETFLOW.

C.1.5 MCG - Evaluation Protocol

In the test set, for each molecule with L ground-truth conformers, we generate $K = 2L$ conformers and evaluate their quality using standard metrics.

Evaluation Metrics. As a conformer C represents an assignment of each atom in the molecular graph to a point in 3D space, it can be viewed as a set of vectors in $\mathbb{R}^{3n}$.To evaluate molecular conformer generation, previous works have employed two key metrics: Average Minimum RMSD (AMR) and Coverage (COV) for both Precision (P) and Recall (R). Given a molecular graph, we generate twice as many conformers as those provided by CREST. Let:

- $\{C_l^*\}_{l=1}^L$ be the set of grounth-truth conformers provided by CREST.
- $\{C_k^*\}_{k=1}^K$ be the set of generated conformers, where $K = 2L$.
- δ be a predefined RMSD threshold for considering a conformer match.

COV-P: Measures the proportion of generated conformers that closely match at least one ground-truth conformer.

$$\text{COV-P} = \frac{1}{K} \left| \{k \in [1,K] \mid \exists l \in [1,L], \text{RMSD}(C_k, C_l^*) < \delta\} \right|$$

AMR-P: Computes the average of the minimum RMSD values between each generated conformer and its closest ground-truth conformer.

$$\text{AMR-P} = \frac{1}{K} \sum_{k=1}^{K} \min_{l=1}^{L} \text{RMSD}(C_k, C_l^*)$$

COV-R: Measures the proportion of ground-truth conformers that have at least one close-enough generated conformer.

$$\text{COV-R} = \frac{1}{L} \left| \{l \in [1,L] \mid \exists k \in [1,K], \text{RMSD}(C_k, C_l^*) < \delta\} \right|$$

AMR-R: Computes the average of the minimum RMSD values between each ground-truth conformer and its closest generated conformer.

$$\text{AMR-R} = \frac{1}{L} \sum_{k=1}^{L} \min_{l=1}^{K} \text{RMSD}(C_k, C_l^*)$$

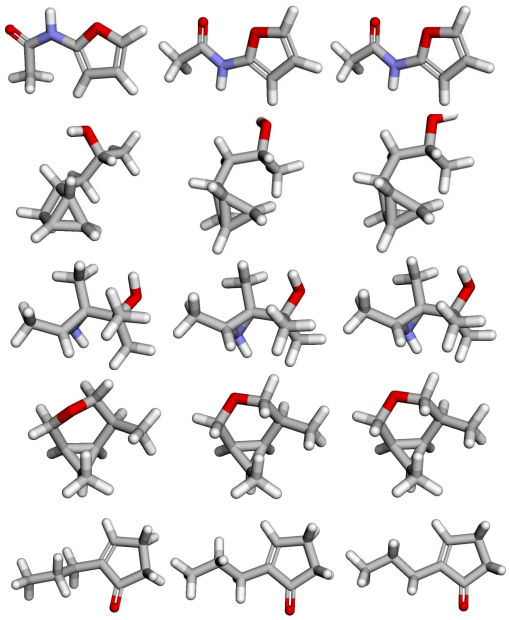


Figure 7: Molecular conformers generated by ETFLOW (left), + [OrbDiff] (center), and ground-truth (right).

C.1.6 MCG - Full results

Table 6: Molecular conformer generation performance on GEOM-QM9. * Reported in the original paper. † Obtained using the published checkpoint. ‡ From our reimplementation trained from scratch.

Models	Recall						Precision					
	Cov@0.1 (↑)		Cov@0.5 (↑)		AMR (↓)		Cov@0.1 (↑)		Cov@0.5 (↑)		AMR (↓)	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GEODIFF	-	-	76.5	**100.0**	0.297	0.229	-	-	50.0	33.5	1.524	0.510
GEOMOL†	28.4	0.0	91.1	**100.0**	0.224	0.194	20.7	0.0	85.8	**100.0**	0.271	0.243
Torsional Diff.†	37.7	25.0	88.4	**100.0**	0.178	0.147	27.6	12.5	84.5	**100.0**	0.221	0.195
MCF†	81.9	**100.0**	94.9	**100.0**	0.103	0.049	78.6	**93.8**	**93.9**	**100.0**	**0.113**	0.055
ETFLOW*	-	-	96.5	**100.0**	0.073	0.047	-	-	94.1	**100.0**	0.098	0.039
ETFLOW†	79.5	**100.0**	93.8	**100.0**	0.096	0.037	74.4	83.3	88.7	**100.0**	0.142	0.066
ETFLOW‡	81.4	**100.0**	94.4	**100.0**	0.092	0.039	74.6	85.5	89.1	**100.0**	0.145	0.064
+ [OrbDiff]	**85.4**	**100.0**	**96.3**	**100.0**	**0.074**	**0.027**	**80.2**	**93.9**	91.9	**100.0**	**0.113**	**0.042**

C.2 Crystal Structure Prediction (CSP)

Crystal Structure Prediction (CSP) is the process of identifying the most stable three-dimensional arrangement of atoms in a crystalline solid, given only its chemical formula. This task lies at the heart of computational materials science, as the resulting crystal structure dictates key physical and chemical properties—including thermodynamic stability, electronic conductivity, and chemical reactivity (Wei et al., 2024; Shao et al., 2021; Kim et al., 2025). Despite its importance, CSP remains a formidable challenge due to the immense combinatorial search space of atomic positions and the presence of complex symmetry constraints that define equivalent configurations (Oganov & Glass, 2006). Efficiently navigating this landscape to discover low-energy, physically plausible structures continues to be a central focus of the field.

We also evaluate our method on a related task introduced by TGDMat (DAS et al., 2025), where crystal structures are generated based on textual descriptions of the desired materials. This setting reflects a more user-centric interface for materials design, where scientists or domain experts can

specify target properties or structural features in natural language. The task includes two levels of textual conditioning: long descriptions, which provide detailed structural and compositional information, and short descriptions, which are more concise and easier to obtain. Supporting text-to-structure generation enables more accessible and flexible workflows in materials discovery, especially in scenarios where precise structural data may be unavailable.

C.2.1 CSP - Dataset

We evaluate our method on two used CSP benchmarks: Perov-5 (Castelli et al., 2012a,b) and MP-20 (Jain et al., 2013). These datasets encompass a broad range of inorganic crystal compositions. Perov-5 comprises 18,928 perovskite structures, which share a common structural motif but vary in elemental composition. In contrast, MP-20 includes 45,231 stable inorganic materials curated from the Materials Project database (Jain et al., NA), offering a more diverse set of crystal systems and chemistries.

C.2.2 CSP - Baselines

For comparison, we use the three strongest baselines from the DiffCSP paper (Jiao et al., 2023): P-cG-SchNet (Gebauer et al., 2022), CDVAE (Xie et al., 2022), and DiffCSP, all with publicly available implementations.

C.2.3 CSP — DiffCSP and TGDMat with Orbit Diffusion

DiffCSP (Jiao et al., 2023) is a diffusion-based framework for crystal structure prediction that jointly models lattice parameters and atomic positions while respecting the fundamental symmetries of crystalline materials. Since TGDMat is built upon DiffCSP, we focus on describing how Orbit Diffusion integrates with DiffCSP; the same integration applies directly to TGDMat.

DiffCSP formulates the task as a *joint diffusion process* with two interconnected components: one for the *lattice* and one for the *atomic coordinates*. The lattice defines the shape and scale of the unit cell, while the coordinates specify atomic positions in fractional units relative to the lattice vectors. To capture the relevant symmetries, the lattice diffusion is $\mathsf{O}(3)$-equivariant (invariant under rotations and reflections), and the coordinate diffusion is both *permutation equivariant* and *periodic translation equivariant*, reflecting atomic indistinguishability and lattice periodicity.

Since the lattice involves only a few parameters, it is a relatively simple subtask. We therefore concentrate on the more challenging component: generating atomic coordinates. To ensure periodic translation equivariance, DiffCSP defines a forward diffusion process based on the kernel $\hat{q}^G(x_t \mid x_0)$, modeled as a *Wrapped Normal* distribution—a periodic analogue of the Gaussian—ensuring the diffusion respects the toroidal geometry of fractional coordinates. Specifically,

$$\hat{q}^G(x_t \mid x_0) \propto \sum_{z \in \mathbb{Z}^d} \exp\left(-\frac{\|x_t - x_0 + z\|^2}{2\sigma_t^2}\right), \tag{31}$$

which defines a valid density on the torus $\mathbb{T}^d$. One can verify that $\hat{q}^G(g \circ x_t \mid g \circ x_0) = \hat{q}^G(x_t \mid x_0)$ for any g in the periodic translation group (Jiao et al., 2023), confirming its equivariance. Consequently, all our theoretical results (e.g., Theorem 1 and Theorem 2) hold when applying Orbit Diffusion to DiffCSP under the periodic translation group.

Orbit Diffusion with non-uniform group sampling. Rather than sampling uniformly from the group G, for the periodic translation group we propose sampling translation elements from a Wrapped Normal distribution. We refer to this variant as `OrbDiff_WN`. Formally,

$$\nu_t(g) \propto \sum_{z_g \in \mathbb{Z}^{d_g}} \exp\left(-\frac{\|m_g + z_g\|^2}{2\sigma_g(t)^2}\right), \tag{32}$$

where m_g is the translation vector corresponding to group element g, and $\sigma_g(t)$ is the time-dependent bandwidth. In our experiments with both DiffCSP and TGDMat, we set $\sigma_g(t) = 2\sigma_t$. To sample from this distribution, we first draw $\epsilon \sim \mathcal{N}(0, I)$ in $\mathbb{R}^3$, and then compute $m_g = \sigma_g(t)\epsilon \mod 1$. We sample 1000 such group elements per step to form the group approximation.

More training details. All models were trained on a single NVIDIA GeForce RTX 4090 GPU. TGDMat was trained for 1,500 epochs, while DiffCSP was trained for 500 epochs. On the MP20 dataset, each epoch took roughly 15 seconds, resulting in total training times of 6.25 hours for TGDMat and 2.08 hours for DiffCSP. On the Perov-5 dataset, each epoch took about 5 seconds, corresponding to 2.08 hours (TGDMat) and 0.69 hours (DiffCSP) of training time.

C.2.4 CSP - Evaluation Protocol

To evaluate crystal structure prediction, we randomly generate one sample for each structure in the test set. We then calculate two metrics: the Match Rate and the average Root Mean Square Distance (RMSD) across the test set. We repeat this procedure three times and report the median values of these metrics for more reliability.

Match rate: The Match Rate is defined as the proportion of predicted structures that successfully match the corresponding ground-truth structures in the test set. Specifically, it is calculated as follows:

$$\text{Match Rate} = \frac{\text{Number of matched structure pairs}}{\text{Total number of test samples}}$$

Following previous works (Jiao et al., 2023; Xie et al., 2022), we use the `StructureMatcher` class from the `pymatgen` library to determine structure matching. The matching process is based on the following criteria:

- Length Tolerance (*ltol*): 0.5 (fractional length tolerance).
- Site Tolerance (*stol*): 0.3 (fraction of the average free length per atom)
- Angle Tolerance (*atol*): 10 (in degrees)

The `StructureMatcher` algorithm aligns the lattice vectors of two structures. If the tolerance criteria are satisfied, the structures are considered matched.

RMSD: For an alignment between lattices of two structures, `StructureMatcher` continues to align atoms to compute the average RMSD. The process is repeated for all possible lattices to find the smallest RMSD. Then the Average RMSD is computed as the average of the smallest RMSD of all matched structure pairs.

$$\text{RMSD} = \frac{1}{N_{\text{matched}}} \sum_{i=1}^{N_{\text{matched}}} \text{RMSD}(\text{generated}_i, \text{ground-truth}_i)$$

Here, N_{matched} is the total number of matched structures. Unmatched structures are excluded from the calculation.

Ideally, we aim for a high Match Rate and a low RMSD. A low Match Rate with a low RMSD is not useful because unmatched samples are effectively treated as having very high RMSD. Thus, RMSD alone cannot fully capture prediction quality. We emphasize that Match Rate is more critical, especially during initial screening, where we prioritize valid structures over perfectly matched ones.

C.2.5 CSP - Full Results and Visualizations

For the TGDMat models (TGDMat (S) and TGDMat (L)), we trained both from scratch, experimenting with and without the proposed loss functions: `[OrbDiff_U]` and `[OrbDiff_WN]`. Meanwhile, the DiffCSP model was trained using our proposed losses, while the baseline models (P-cG-SchNet, CDVAE, and DiffCSP) rely on the results reported in the original DiffCSP paper. Quantitative results are summarized in Tables 7 and 8 with qualitative comparisons shown in Figures 8 and 9.

Table 7: Text-guided CSP with TGDMat.

Method	Perov-5		MP-20	
	Match (↑)	RMSE (↓)	Match (↑)	RMSE (↓)
TGDMat (S)	59.39	0.066	59.90	0.078
+ [OrbDiff_U]	63.51	0.062	56.50	0.085
+ [OrbDiff_WN]	**65.57**	**0.054**	**61.29**	**0.072**
TGDMat (L)	95.17	0.013	61.91	0.081
+ [OrbDiff_U]	95.88	**0.012**	65.94	**0.069**
+ [OrbDiff_WN]	**95.98**	**0.012**	**66.74**	**0.069**

Table 8: Crystal Structure Prediction (CSP).

Method	Perov-5		MP-20	
	Match (↑)	RMSE (↓)	Match (↑)	RMSE (↓)
P-cG-SchNet	48.22	0.418	15.39	0.376
CDVAE	45.31	0.114	33.90	0.105
DiffCSP	52.02	0.076	51.49	0.063
+ [OrbDiff_U]	52.29	0.078	54.47	0.054
+ [OrbDiff_WN]	**52.39**	**0.069**	**55.70**	**0.053**

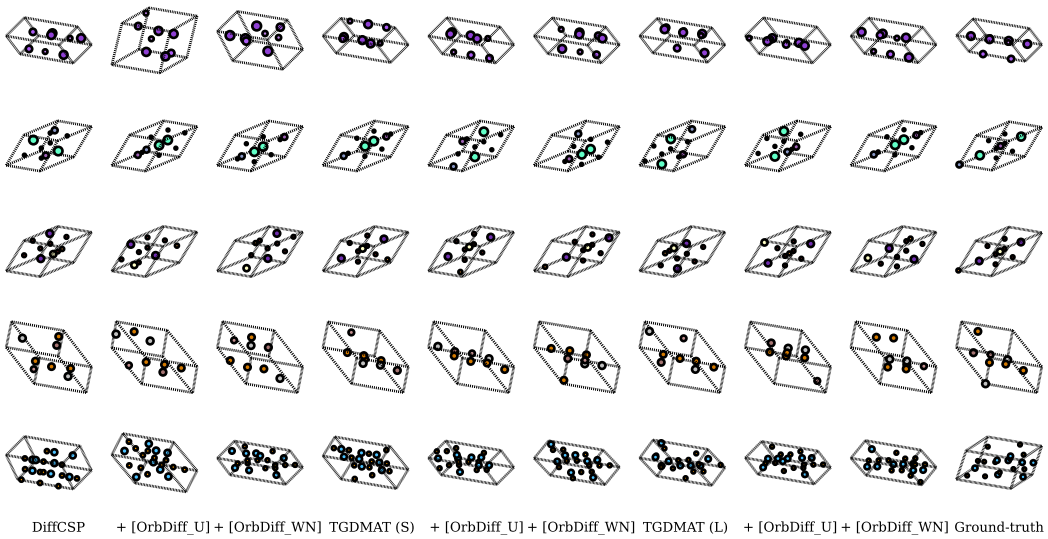


Figure 8: Qualitative comparison of Crystal Structure Predictions by 9 models, including DiffCSP, TGDMAT short and TGDMAT long with baselines, OrbDiff_U, and OrbDiff_WN against ground-truth samples on randomly selected samples from MP-20 dataset.

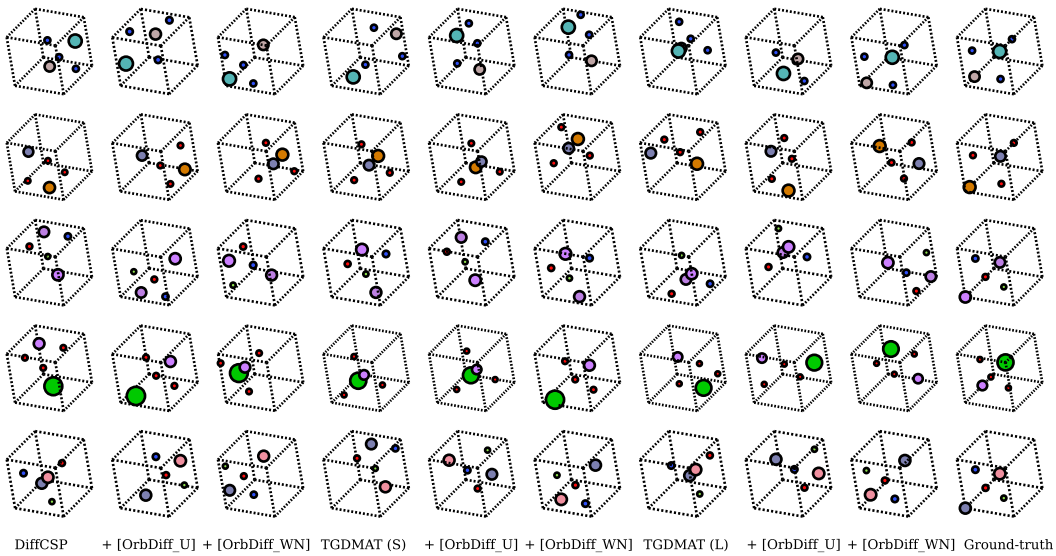


Figure 9: Qualitative comparison of Crystal Structure Predictions by 9 models, including DiffCSP, TGDMat (S) and TGDmat (L) with baselines, OrbDiff_U, and OrbDiff_WN against ground-truth samples on randomly selected samples from Perov-5 dataset.

C.3 Protein Structure Generation (PSG) with Non-Equivariant Denoiser PROTEINA

Protein structure generation focuses on sampling physically valid 3D conformations of proteins by learning a probabilistic distribution over either atomistic or coarse-grained representations. This generative task plays a crucial role in de novo protein design and has broad implications for understanding protein folding and function (Jing et al., 2023; Wu et al., 2022; Watson et al., 2023; Huguet et al., 2024; Jing et al., 2023). Unlike traditional structure prediction, which aims to infer a single or most likely conformation, generative models must capture the full structural manifold—accounting for the inherent $\mathsf{SO}(3)$ rotational symmetry of protein backbones, maintaining biochemical realism, and respecting physical constraints such as bond lengths and steric clashes (Gaujac et al., 2024; Geffner et al., 2025). Effectively modeling these aspects ensures that generated structures are not only diverse but also biologically and physically plausible.

C.3.1 PSG - Baselines

We list the baselines used by the work of (Geffner et al., 2025) as follows:

- FrameDiff (Yim et al., 2023b)
- FoldFlow (Bose et al., 2024)
- FrameFlow (Yim et al., 2023a)
- ESM3 (Hayes et al., 2025)
- Chroma (Ingraham et al., 2023)
- RFDiffusion (Watson et al., 2023)
- Proteus (Wang et al., 2024a)
- Genie2 (Lin et al., 2024)

PROTEINA. PROTEINA is a large-scale, flow-based model for generating protein backbones, built on a scalable transformer architecture. Unlike equivariant models, it does not enforce equivariance, granting more architectural flexibility. This choice allows the use of powerful transformer networks with hundreds of millions of parameters, enabling PROTEINA to effectively learn from large datasets. As a result, it excels at modeling complex protein structures by balancing expressiveness and computational scalability.

There are three variants of PROTEINA:
(i) $\mathcal{M}_{\text{FS}}$, a 200M-parameter transformer with an additional 15M parameters in triangle layers, trained on the **Foldseek AFDB clusters** dataset $\mathcal{D}_{\text{FS}}$, which includes 555,318 structures of lengths 32–256;
(ii) $\mathcal{M}_{\text{FS}}^{\text{no-tri}}$, a simplified version without triangle layers, also with 200M parameters and trained on the same dataset;
(iii) $\mathcal{M}_{\text{21M}}$, a 400M-parameter transformer with 15M triangle parameters, trained on a **high-quality filtered AFDB subset** $\mathcal{D}_{21M}$, comprising approximately 21M structures. $\mathcal{M}_{\text{21M}}$ represents the current state of the art in designability modeling.

C.3.2 PROTEINA with Orbit Diffusion

We apply Orbit Diffusion to the simplest variant of PROTEINA, namely $\mathcal{M}_{\text{FS}}^{\text{no-tri}}$, by fine-tuning the public checkpoint released by Geffner et al. (2025), since we do not have access to the extensive compute resources used for the original training. Our fine-tuning setup uses 4 A100 GPUs for 24 hours on the same dataset, whereas the original training employed 96 GPUs.

The key symmetry group in protein structure generation is $\mathsf{SO}(3)$, which represents 3D rotations. To exploit this symmetry using `OrbDiff`, we apply 10,000 uniformly sampled random rotations to each training sample as part of our Rao-Blackwell estimator, enhancing both the efficiency and stability of the flow matching process.

The only change we introduce to the original model is replacing the conditional flow matching loss with our proposed flow matching objective (see Appendix C.1.3).

For a fair comparison, we also fine-tune $\mathcal{M}_{\text{FS}}^{\text{no-tri}}$ (+ [finetune]) using the original loss under the same computational budget, with a batch size of 8 and 32 gradient accumulation steps.

C.3.3 PSG - Evaluation Protocol

To evaluate the quality of our generated protein backbones, we rely on three widely used metrics: designability, diversity, and novelty. Following the protocol established by Geffner et al. (2025), we generate 500 samples total—100 for each length in {50, 100, 150, 200, 250}—and compute all metrics on this dataset. Among these, **designability is the most critical metric**, as it directly reflects the biological feasibility of the generated structures and serves as the foundation for the other two metrics.

Designability. Designability measures whether a backbone structure can realistically be encoded by an amino acid sequence. For each generated backbone, we produce 8 candidate sequences using ProteinMPNN (Dauparas et al., 2022) with a sampling temperature of 0.1. These sequences are then folded using ESMFold (Lin et al., 2023), and the root mean square deviation (RMSD) is calculated between each predicted structure and the original backbone. A backbone is deemed designable if at least one sequence folds with an RMSD below 2Å, where this minimum RMSD is known as the self-consistency RMSD (scRMSD).

Since **diversity and novelty are computed only on designable samples**, accurate assessment of designability is essential for interpreting the other metrics meaningfully.

The designability score for a model is reported as the fraction of samples deemed designable. Additionally, we report the average scRMSD across all samples, allowing for a more nuanced comparison between our model and existing baselines.

Diversity. We evaluate diversity among the designable samples using two approaches. First, we compute the average pairwise TM-score within each length group (50, 100, 150, 200, and 250) as a measure of structural variation. Lower average TM-scores indicate greater diversity.

Second, we calculate diversity (cluster) by grouping designable samples into clusters based on a TM-score threshold of 0.5. Each cluster contains samples with pairwise TM-scores above this threshold. Diversity (cluster) is then defined as the ratio of the total number of clusters to the number of designable samples. A higher ratio reflects a larger number of distinct structural groups relative to the sample size, signaling increased diversity.

Novelty. Novelty measures how structurally distinct the designable samples are compared to known protein structures. For each designable backbone, we compute the TM-score to its closest match in two reference sets: the Protein Data Bank (PDB) and the $\mathcal{D}_{\text{FS}}$ dataset used for training. The average of these maximum TM-scores across all designable samples is reported as the novelty score. Lower values indicate the model generates structures that are more novel relative to both established experimental data and the training distribution.

C.3.4 PSG - Full results

Table 9: **Protein Structure Generation.** Full comparison with baseline models; all baseline results are taken from (Geffner et al., 2025). "+ [finetune]" indicates $\mathcal{M}_{FS}^{\text{no-tri}}$ finetuned with the original loss, while "+ [`OrbDiff`]" denotes finetuning with `OrbDiff`.

Model	Designability		Diversity		Novelty	
	Fraction (↑)	scRMSD (↓)	Cluster (↑)	TM-score (↓)	PDB (↓)	AFDB (↓)
FrameDiff	65.4	-	0.39	0.40	0.73	0.75
FoldFlow (base)	96.6	-	0.42	0.75	0.75	0.77
FoldFlow (stoc.)	97.0	-	0.61	0.38	0.62	0.68
FoldFlow (OT)	97.2	-	0.37	0.41	0.71	0.75
FrameFlow	88.6	-	0.59	0.34	0.79	0.80
ESM3	22.0	-	0.52	0.57	0.70	0.75
Chroma	78.8	-	0.42	0.43	0.77	0.76
RFDiffusion	94.4	-	0.46	0.34	0.79	0.80
Proteus	94.4	-	0.42	0.43	0.77	0.80
Genie2	95.2	-	0.59	0.38	**0.63**	0.69
$\mathcal{M}_{21\text{M}}$	**99.0**	**0.72**	0.30	0.39	0.81	0.84
$\mathcal{M}_{\text{FS}}^{\text{no-tri}}$	93.8	1.04	**0.62**	**0.36**	**0.69**	**0.76**
+ [finetune]	93.8	1.00	0.54	0.37	0.74	0.83
+ [`OrbDiff`]	**95.6**	**0.93**	0.52	0.37	0.74	0.83

The state-of-the-art model $\mathcal{M}_{21\mathrm{M}}$, with 400M parameters and trained on a large, high-quality dataset, achieves the highest designability (99.0%) and lowest scRMSD (0.72), reflecting its strong reconstruction capability. However, this comes at the expense of diversity and novelty: it exhibits the lowest diversity score (Cluster: 0.30) and higher novelty metrics (PDB: 0.81, AFDB: 0.84), indicating reduced structural variety and generalization.

In comparison, our base model $\mathcal{M}_{\mathrm{FS}}^{\text{no-tri}}$ already achieves competitive performance (designability: 93.8%, Cluster: 0.62, PDB: 0.69), and naive finetuning with data augmentation ("+ [finetune]") fails to improve designability or diversity. Notably, our method ("+ [`OrbDiff`]") improves designability to 95.6% and reduces scRMSD to 0.93, while preserving competitive diversity (Cluster: 0.52) and novelty (PDB: 0.74). This highlights that `OrbDiff` is an effective finetuning strategy to enhance functional accuracy without fully sacrificing structural diversity—achieving a better trade-off than both naive finetuning and heavily overparameterized models.

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 clearly state the paper's main claims and contributions, which align with the theoretical and experimental results presented.

   Guidelines:

   - The answer NA 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 NA 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: We have discussed about our limitation in the Section 6.

   Guidelines:

   - The answer NA 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: We provided comprehensive proof of our theoretical results in the Appendix section.

Guidelines:

- The answer NA 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: We provide detailed descriptions of the experimental setup, model architecture, hyperparameters, training procedure, and evaluation protocols.

   Guidelines:

   - The answer NA 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: [NA]

Justification: The paper reuses existing benchmark datasets and builds on top of previous open-source models.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify relevant training and testing details, including data splits, hyperparameters, and optimizer choices. These details align with those used in the reused code repositories for the baseline methods.

   Guidelines:

   - The answer NA 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: [No]

   Justification: Not all experiments include error bars or statistical significance measures due to the high computational cost of training some models, limiting repeated runs.

   Guidelines:

   - The answer NA 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: The paper details the compute resources used of all experiments in Appendix C.

   Guidelines:

   - The answer NA 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: The research fully complies with the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer NA 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: [NA]

    Justification: This work is foundational research without direct societal impact or applications that raise ethical concerns.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not release data or models with high risk for misuse.

    Guidelines:

    - The answer NA 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: This work reuses existing assets with MIT license.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: This paper does not introduce or release any new assets.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: The paper does not involve crowdsourcing or any research with human subjects.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: This research does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used in any core or non-standard part of the method. Their use, if any, was limited to minor writing edits and did not impact the research.

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.