SYMMETRY-PRESERVING DIFFUSION MODELS VIA TARGET SYMMETRIZATION

Vinh Tong[1,2] **Yun Ye**[1], **Trung-Dung Hoang**[4], **Anji Liu**[1,3], **Guy Van den Broeck**[3], **Mathias Niepert**[1,2]

[1]University of Stuttgart, [2]IMPRS-IS, [3]UCLA, [4]University of Bern
vinh.tong@ki.uni-stuttgart.de

ABSTRACT

Diffusion models are powerful tools for capturing complex distributions, but modeling data with inherent symmetries, such as molecular structures, remains challenging. Equivariant denoisers are commonly used to address this, but they introduce architectural complexity and optimization challenges, including noisy gradients and convergence issues. We propose a novel approach that enforces equivariance through a symmetrized loss function, which applies a time-dependent weighted averaging operation over group actions to the model's prediction target. This ensures equivariance without explicit architectural constraints and reduces gradient variance, leading to more stable and efficient optimization. Our method uses Monte Carlo sampling to estimate the average, incurring minimal computational overhead. We provide theoretical guarantees of equivariance for the minimizer of our loss function and demonstrate its effectiveness on synthetic datasets and the molecular conformation generation task using the GEOM-QM9 dataset. Experiments show improved sample quality compared to existing methods, highlighting the potential of our approach to enhance the scalability and practicality of equivariant diffusion models in generative tasks.

1 INTRODUCTION

Diffusion models have proven effective in capturing complex distributions (Ho et al., 2020; Song et al., 2020a;b; Karras et al., 2022), with notable applications in molecular and protein generation (Zhang et al., 2023; Vignac et al., 2023; Anand & Achim, 2022). Many physical systems, such as molecular structures, exhibit inherent symmetries. For instance, a molecule's spatial configuration remains invariant under rotations in three-dimensional space (Hoogeboom et al., 2022; Jing et al., 2022).

Symmetries in data are naturally described by G-*invariant* distributions, which remain unchanged under transformations from a symmetry group G (Chen et al., 2024; Köhler et al., 2020). Our goal is to model a G-*invariant* distribution $q(x_0)$ using diffusion models. In practice, however, we only have access to a finite dataset $D = \{x_0^i\}_{i=1}^N$, which defines the empirical distribution $\hat{q}(x) := \frac{1}{N}\sum_{i=1}^N \delta(x - x_0^i)$, where δ is the Dirac delta function. Unfortunately, $\hat{q}(x_0)$ is not necessarily G-*invariant*, as the dataset D may not be closed under the group action $g \in G$. Consequently, fitting a model $p_\theta(x_0)$ to $\hat{q}(x_0)$ without enforcing equivariant constraints may result in poor generalization. We refer to this approach as *Non-Symmetrized Diffusion* (NSD).

While some NSD models achieve strong performance without explicit equivariant constraints, they typically rely on datasets with a canonical coordinate system (Wang et al., 2024) or large datasets that mitigate the need for augmentation (Abramson et al., 2024). Nevertheless, incorporating equivariant inductive biases into diffusion models remains a preferred approach, as it better captures the underlying symmetries of the data. Two common strategies for achieving this are data augmentation and the design of equivariant denoisers.

Data augmentation involves constructing a symmetrized dataset $D^G = \{g \circ x_0^i \mid g \in G, x_0^i \in D\}$, which yields a G-*invariant* empirical distribution closer to the true distribution $q(x_0)$ (Chen et al., 2024). However, this approach can be computationally expensive, especially when the group G is large or infinite. We refer to this method as *Symmetrized Diffusion* (SD).

Alternatively, one can train diffusion models on the original dataset using an equivariant denoiser (Yim et al., 2023; Hoogeboom et al., 2022; Jing et al., 2022; Anand & Achim, 2022). Since the diffusion sampling process involves repeated denoising steps, an equivariant denoiser ensures that an initially invariant prior distribution remains invariant throughout sampling. This approach has been successfully applied to tasks such as 3D molecular generation (Hoogeboom et al., 2022; Hassan et al., 2024; Guan et al., 2023), protein backbone generation (Bose et al., 2023; Jing et al., 2023), and point cloud generation (Luo & Hu, 2021; Mikuni et al., 2023; Wu et al., 2023).

Formally, enforcing equivariance in diffusion models can be framed as a constrained optimization problem. However, this introduces challenges such as increased architectural complexity, higher computational costs (Wang et al., 2024; Brehmer et al., 2024), and convergence issues due to noisy gradients (Abbe & Boix-Adserà, 2022). In this work, we demonstrate that the loss function in such approaches is inherently noisy, exacerbating these challenges.

To address these issues, we propose a novel approach that explicitly enforces equivariance through a carefully designed loss function. Our method symmetrizes the model's prediction target via a time-dependent weighted averaging operation over group actions applied to data samples. This ensures that the model respects the desired symmetries without requiring explicit architectural constraints. We estimate the average using Monte Carlo sampling, which incurs minimal computational overhead as it avoids additional neural network passes. The resulting loss function not only preserves equivariance but also reduces gradient estimator variance, improving both optimization stability and efficiency.

We provide a theoretical analysis of our method, including a guarantee of equivariance for the minimizer of the proposed loss function. Our experiments on synthetic datasets and the molecular conformation generation task (Hawkins, 2017), using the GEOM-QM9 dataset (Axelrod & Gomez-Bombarelli, 2022), demonstrate that our approach improves sample quality compared to existing methods. These results underscore the potential of our method to enhance the scalability and practicality of equivariant diffusion models in generative tasks.

2 Background

This section introduces key mathematical concepts, starting with groups and group actions which ensures symmetry in generative models. We then discuss diffusion models and equivariant diffusion models for generating structured data that respects these symmetries.

2.1 Group and Group Action

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. Groups are fundamental in describing symmetries and transformations in mathematics and physics. A *group action* defines how a group G operates on a set Ω. Formally, it is a bijective map $\circ : G \times \Omega \to \Omega$ satisfying the identity and compatibility properties, ensuring the action is well-defined and consistent. Here, Ω is a set on which the group acts, such as a set of data points or geometric objects.

A function $f : \Omega \to \mathbb{R}$ is said to be *G-invariant* if for all $g \in G$ and $x \in \Omega$, the function satisfies:

$$f(g \circ x) = f(x).$$

This means that the function value does not change under the action of any element of the group G.

On the other hand, a function $f : \Omega \to \Omega$ is said to be *G-equivariant* if for all $g \in G$, $x \in \Omega$, the following condition holds:

$$f(g \circ x) = g \circ f(x),$$

where the action of G on Ω is understood in an appropriate way, ensuring that the group action commutes with the transformation of f. In this paper, we only consider *linear isometry groups*, i.e., any mapping $g : \Omega \to \Omega$ can be written as $g(x) \mapsto A_g x$ for some unitary matrix $A_g \in \mathbb{R}^{d \times d}$. Examples of such isometry groups include the permutation group S_n, which consists of all possible permutations of n elements, the orthogonal group $\mathsf{O}(d)$, which includes rotations and reflections in d-dimensional space, and the special orthogonal group $\mathsf{SO}(d)$, which consists of rotations only.

Invariant Distribution A probability distribution $p(x)$ defined on a set Ω is said to be *G-invariant* under the action of a group G if for all $g \in G$ and measurable subsets $A \subset \Omega$, the following holds:

$$p(g \circ x \in A) = p(x \in A),$$

where $g \circ x$ denotes the action of $g \in G$ on $x \in \Omega$. Intuitively, this means that the distribution remains unchanged under the transformations induced by the group.

Equivariant Conditional Distribution A conditional distribution $p(y \mid x)$, where $x \in \Omega$ and $y \in \mathcal{Y}$, is said to be *G-equivariant* if for all $g \in G$, the following condition holds:

$$p(g \circ y \mid g \circ x) = p(y \mid x).$$

This property ensures that the conditional distribution transforms consistently with the group action, preserving the symmetry in the relationship between x and y. An example of an equivariant conditional distribution is the Gaussian kernel $q(x \mid y) = \mathcal{N}(x; y, \sigma^2 I)$, which satisfies $q(g \circ x \mid g \circ y) = q(x \mid y)$ for all isometry groups.

Group Symmetrization. We use group symmetrization operators (Chen et al., 2024; Birrell et al., 2022), which play a crucial role in developing the theoretical framework for our approach. Let S_G be the symmetrization operator under group G, transforming any distribution $p(x)$ into a *G-symmetrized* distribution:

$$S_G[p](x) := \int_G p(g \circ x) \mathrm{d}\mu_G(g),$$

where μ_G is the unique Haar probability measure on G. Importantly, this *G-symmetrized* distribution is a *G-invariant* distribution (Birrell et al., 2022). Group symmetrization can be useful to study the theoretical properties of equivariant distributions (Chen et al., 2024; Birrell et al., 2022) or to design equivariant neural networks (Puny et al., 2022).

2.2 Diffusion Models and Equivariant Diffusion Models

Diffusion models (Ho et al., 2020; Song et al., 2020b) are a class of generative models that construct complex data distributions by iteratively transforming simple noise distributions through a learned denoising process. The core idea is to define a forward diffusion process that incrementally corrupts the data and a reverse process that learns to reconstruct the original data. 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., 2020b) 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 $t = T$, $q(x_T) = \mathcal{N}(0, I)$.

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:

$$\mathcal{L} = \mathbb{E}_{t \sim \mathcal{U}(0,T), (x_0, x_t) \sim q(x_0, x_t)} \left[\omega(t) \|\phi_\theta(x_t, t) - x_0\|^2 \right], \tag{1}$$

where $\omega(t)$ is a time-dependent loss weight. 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 design of the reverse process offers significant flexibility. It can involve solving an ODE or SDE numerically (Song et al., 2020b; Karras et al., 2022; Lu et al., 2022; Tong et al., 2024), or it may employ ancestral sampling methods (Ho et al., 2020).

Equivariant Diffusion Models. Many generative tasks involve data with inherent symmetries, such as rotation, reflection, or permutation invariance, which must be preserved. Equivariant diffusion models extend the standard diffusion framework by enforcing equivariance of the neural network ϕ_θ. A model is said to be *G-equivariant* if the denoiser ϕ_θ satisfies $\phi_\theta(g \circ x_t, t) = g \circ \phi_\theta(x_t, t)$, where $g \in G$ is a symmetry transformation from the group G. Diffusion models with *G-equivariant* denoisers can model *G-invariant* distributions (Chen et al., 2024). Intuitively, the Gaussian prior

distribution is inherently G-*invariant*, and the sampling process of a diffusion model can be viewed as a composition of denoising steps, each being an equivariant function. As a result, the entire denoising process becomes a G-*equivariant* function, which maps a G-*invariant* distribution to another G-*invariant* distribution, thereby preserving invariance throughout the sampling process.

These models are particularly suitable for applications in structural biology (Corso et al., 2024; Yim et al., 2023; Schneuing et al., 2024a; Igashov et al., 2024) and molecular modeling (Hoogeboom et al., 2022; Guan et al., 2023; Le et al., 2023), where data resides in geometric domains governed by the linear isometry group. By ensuring that the generative process respects these symmetries, equivariant diffusion models produce results that are more physically meaningful.

3 METHOD

We propose a new loss function that attains the same minimizer as *Symmetrized Diffusion* (SD) without requiring explicit constraints on the denoiser or an increase in training dataset size. While our approach does not rely on these techniques, they can still be incorporated if desired. In our experiments, for instance, we also use a G-equivariant denoiser. Furthermore, we show that our loss function exhibits lower gradient variance than existing approaches.

To provide a clearer foundation for our analysis, we first present a theoretical observation that establishes the relationship between the forward diffusion processes of the original empirical distribution and its symmetrized counterpart. We assume that our dataset consists of unique structures, meaning no two samples are equivalent under group transformations. Consider two forward diffusion processes: one applied to the original empirical distribution $\hat{q}(x_0)$, generating time-dependent marginals $\hat{q}_t(x_t)$, and the other applied to the symmetrized distribution $\hat{q}^G(x_0) = S_G[q](x_0)$, generating $\hat{q}_t^G(x_t)$. We show that $\hat{q}_t^G(x_t)$ is the symmetrized version of $\hat{q}_t(x_t)$, i.e., $\hat{q}_t^G(x_t) = S_G[\hat{q}_t](x_t)$. The formal lemma and proof are provided in Appendix B.1, and an illustration is shown in Figure 1. Our objective is to

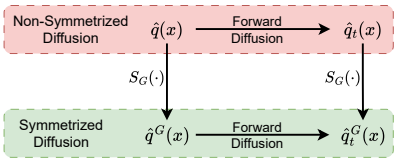


Figure 1: The forward diffusion process applied to the symmetrized distribution $\hat{q}^G(x_0)$ yields the symmetrized time-dependent marginal distributions $\hat{q}_t^G(x_t)$ of the original process $\hat{q}_t(x_t)$.

design a loss function with the same minimizer as SD without explicitly constraining the denoiser or relying on data augmentation.

Analyzing the Minimizer of Symmetrized Diffusion. We examine the minimizer of the empirical diffusion loss at time t:

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

The minimizer of the empirical loss function is the point where its derivative with respect to ϕ becomes zero. For each x_t, the gradient is given by:

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

Thus, the minimizer ϕ^* of Equation (2) satisfies:

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

which is *G-equivariant* (proof in Appendix B.2), while the Non-symmetrized Diffusion's minimizer is not necessarily *G-equivariant* (see Appendix B.3 for the proof).

Although *Symmetrized Diffusion* has an equivariant minimizer, data augmentation can significantly increase the dataset size. This, in turn, requires greater computational resources.

A common alternative is to design the denoiser architecture to naturally handle data symmetries. This often involves using specialized layers where the weight matrices and activation functions follow specific constraints (Hoogeboom et al., 2022; Anand & Achim, 2022; Schneuing et al., 2024b).

Constrained Equivariant Diffusion Models Enforcing ϕ to be G-equivariant while minimizing the diffusion loss involves solving the following constrained optimization problem:

$$\begin{aligned} &\text{Minimize} \quad \mathcal{L}_t(\phi) = \omega(t)\mathbb{E}_{(x_0,x_t)\sim\hat{q}_t(x_0,x_t)}[\|\phi(x_t,t) - x_0\|^2] \\ &\text{Subject to} \quad \phi(g \circ x_t, t) = g \circ \phi(x_t, t) \quad \forall x_t \in \Omega \text{ and } \forall g \in G. \end{aligned} \tag{3}$$

This formulation, where equivariance constraints are explicitly enforced, has gained significant attention in recent research (Yim et al., 2023; Hoogeboom et al., 2022). Recent theoretical work (Chen et al., 2024) demonstrates that *Constrained Equivariant Diffusion Models* and *Symmetrized Diffusion* share the same minimizer.

In practice, this optimization is challenging and often requires complex design choices to enforce equivariance, which can restrict the class of minimizers and complicate training (Elhag et al., 2024). Furthermore, since the minimizer of $\mathcal{L}_t(\phi)$ is not naturally equivariant (see proof in Appendix B.3), the equivariant constraint must actively guide the optimization process, adding to its difficulty. Additionally, we will show that the Monte Carlo estimation of the loss function exhibits high variance, further exacerbating the optimization challenge (Abbe & Boix-Adserà, 2022).

Proposed Loss Function. To address the challenges of *Symmetrized Diffusion* and *Constrained Equivariant Diffusion Models*, we propose a novel loss function:

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

where

$$\phi_G^*(x_t,t) = \int_\Omega x_0 \hat{q}_t^G(x_0 \mid x_t)\mathrm{d}x_0.$$

Rather than denoising x_t directly to the target x_0, this approach encourages $\phi(x_t,t)$ to align with $\phi_G^*(x_t,t)$, which is already a *G-equivariant* function (proof in Appendix B.2). Notably, the minimizer of this loss function is $\phi_G^*(x_t,t)$, eliminating the need for data augmentation or explicitly enforcing equivariance in the network design.

Although computing $\phi_G^*(x_t,t)$ analytically is intractable, we can approximate it using the Monte Carlo method outlined in Algorithm 1. To understand how this approximation is derived, we decompose the integral over Ω into a double integral over the quotient space Ω/G and the group G, since our dataset consists of unique structures:

$$\begin{aligned} \phi_G^*(x_t,t) &= \int_{\Omega/G}\int_G (g\circ x_0)\hat{q}_t^G(x_0 \mid x_t)\mathrm{d}\mu_G(g)\mathrm{d}x_0 \\ &= \sum_{x_0\in D}\int_G (g\circ x_0)\hat{q}_t^G(x_0 \mid x_t)\mathrm{d}\mu_G(g). \end{aligned}$$

We approximate this integral by first sampling x_0 from the dataset D and then averaging over transformations g drawn from the group G.

While the proposed loss inherently aligns with the minimizer of the constrained problem, practical optimization may still introduce potential deviations. Therefore, our method does not rule out the use of the equivariant constraint if one chooses to apply it. There are two key benefits to using our loss function in conjunction with the equivariant constraint.

First, the minimizer of our loss function already satisfies the equivariant constraint, meaning that the role of the constraint is no longer solely to guide the minimizer to the correct equivariant functional class. This significantly simplifies the optimization problem. Second, we will show that our loss function exhibits low variance, which further aids in stabilizing and accelerating the optimization process.

Practical Implementation. We prove in Appendix B.4 that $\phi_G^*(x_t, t)$ can be simplified to the following form:

$$\phi_G^*(x_t, t) = C \sum_{x_0 \in D} \int_G (g \circ x_0) \exp\left(\frac{-\|x_t - \alpha_t(g \circ x_0)\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu_G(g),$$

where C is a normalized factor such that:

$$C \sum_{x_0 \in D} \int_G \exp\left(\frac{-\|x_t - \alpha_t(g \circ x_0)\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu_G(g) = 1.$$

The function $\phi_G^*(x_t, t)$ represents a normalized expectation over transformations of data samples $x_0 \in D$ under the action of the group G. Specifically, it computes a weighted average of the transformed inputs $g \circ x_0$, with weights determined by a Gaussian kernel that measures the similarity between the noisy sample x_t and the scaled transformed input $\alpha_t(g \circ x_0)$. We can approximate $\phi_G^*(x_t, t)$ using Monte Carlo sampling to estimate the group integral, as detailed in Algorithm 1.

Algorithm 1 Approximation of $\phi_G^*(x_t, t)$

1: Draw M samples $x_0^m \sim \hat{q}(x_0)$, K group elements $g_k \sim \mu_G$
2: **for** each x_0^m, g_k **do**
3: $\quad x_0^{mk} \leftarrow g_k \circ x_0^m, \quad R_{mk} \leftarrow \frac{-\|x_t - \alpha_t x_0^{mk}\|^2}{2\sigma_t^2}$
4: **end for**
5: $w_{mk} \leftarrow \frac{\exp(R_{mk})}{\sum_{j=1}^{MK} \exp(R_j)}, \quad \phi_G^*(x_t, t) \leftarrow \sum_{j=1}^{MK} w_j x_0^j$

A key advantage of the proposed loss function is its significantly lower gradient estimator variance when used with a *G-equivariant* denoiser, compared to the constrained optimization objective in Eq. (3). This reduction arises from the symmetrization in $\phi_G^*(x_t, t)$, which averages out noise over group transformations. Formally:

Variance Reduction of the Proposed Loss

Theorem 1. *Let $\mathcal{L}_t(\phi)$ and $\bar{\mathcal{L}}_t^G(\phi)$ be the Monte Carlo estimators corresponding to Eq.* (3) *and Eq.* (4)*, respectively, where $\phi(x_t, t)$ is G-*equivariant*. Then:*

$$\mathrm{Var}(\nabla_\phi \mathcal{L}_t(\phi)) \geq \mathrm{Var}(\nabla_\phi \bar{\mathcal{L}}_t^G(\phi)). \tag{5}$$

A full proof is provided in Appendix B.5.

4 Experimental Results

4.1 Toy Experiment

We evaluate our loss using the most common design of equivariant diffusion models—*Non-Symmetrized Diffusion* with an Equivariant Denoiser (NSD+Enet)—on two toy examples. Each example consists of a single data sample and illustrates a scenario where the empirical distribution is highly asymmetric.

Reflection group in 1D. In this toy experiment, we create a dataset containing a single 1D sample with value 1. The equivariant group is the reflection group (i.e., $g \cdot x = \pm x$). We train a denoiser $D_\theta(x_t, t) = f_\theta(x_t, t) - f_\theta(-x_t, t)$, which is a *G-equivariant* function. The function f_θ is a 3-layer multi-layer perceptron (MLP), with hidden sizes of 2, 10, and 5, and Tanh activation functions.

We train the models with a batch size of 100, a learning rate of 0.001, and a total of 100,000 iterations. During training, we log the loss and periodically evaluate the model. After training, we generate

100,000 samples using 100 DDIM (Song et al., 2020a) steps and compute the root mean square deviation (RMSD) between each sample and its closest target value in $\{-1, 1\}$. Additionally, we calculate the Wasserstein-2 (W2) distance between the generated samples and the target distribution. We repeat the evaluation three times with different random seeds and report the mean and standard deviation in Table 1.

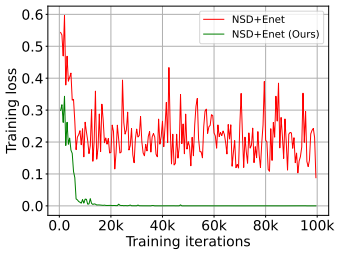


Figure 2: Training Loss.

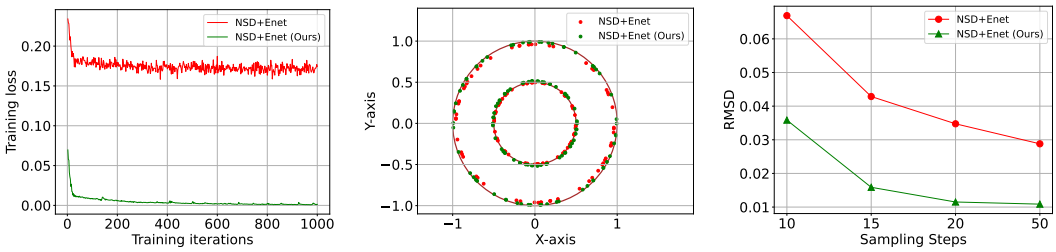
Models	RMSD ($\times 10^{-3}$)	W2 Distance ($\times 10^{-2}$)
NSD+Enet	8.3923 $\pm$ 1.8076	0.0139 $\pm$ 0.0032
NSD+Enet (Ours)	**2.0486** $\pm$ **1.4348**	**0.0017** $\pm$ **0.0005**

Table 1: 1D Toy Experiment Results. We report RMSD between the generated samples and their closest target in $\{-1, 1\}$. We also report the W2 Distance between the generated samples and the ground-truth distribution.

Our results demonstrate that our loss function improves both training stability and model performance across different settings. Figure 2 shows that our loss function stabilizes training, which leads to faster and more stable convergence. In contrast, training without our loss function exhibits high variance and fails to converge properly, as seen in fluctuating training loss curves.

Our model achieves significantly better results than the baseline. Specifically, our loss achieves 4 times better RMSD and 8 times better W2 Distance than the original diffusion loss. This demonstrates that our approach generates samples much closer to the target values while aligning more accurately with the ground truth distribution.

Rotation and Permutation Groups in 2D. In this experiment, the model is tasked with generating a point cloud of 4 points arranged in a straight line, ensuring equivariance to both rotation and permutation groups. Similar to the 1D experiment, the training dataset consists of a single sample. This sample comprises 4 points in 2D space with coordinates $(-1.0, 0.0)$, $(-0.5, 0.0)$, $(0.5, 0.0)$, and $(1.0, 0.0)$, respectively. We use EGNN (Satorras et al., 2021) as the denoiser architecture, constructing a fully connected graph from the 4 points and using their coordinates as node features.



(a) Training Loss (first 1k iterations) (b) Qualitative Visualization (c) RMSD - lower is better

Figure 3: **2D toy experiment.** (a) Model's convergence during training. (b) Visualization of 100 samples generated by each model, where accurate samples have points lying within the brown circles. (c) Comparison of RMSD between model samples and the ground truth, showing the average RMSD over 1000 samples with sampling steps ranging from 10 to 50.

Figure 3 demonstrates the effectiveness of our proposed loss function when applied to NSD+Enet. Similar to the 1D experiment, Figure 3a shows that our loss function results in a significantly more stable training loss curve compared to the original loss function. Figure 3c shows that using our loss function leads to much better sample quality and requires fewer sampling steps to achieve high-quality samples. For instance, NSD+Enet (Ours) with 15 sampling steps achieves an RMSD of around 0.15, outperforming NSD+Enet with 50 sampling steps, which results in an RMSD of 0.29.

4.2 MOLECULAR CONFORMER PREDICTION

To evaluate our method in real-world scenarios, we apply it to molecular conformer generation, a crucial task in drug discovery. Given a molecule's 2D molecular graph $G = (V, E)$, the goal is to predict its 3D conformations. A conformer C represents a mapping of each atom in V to a point in 3D space and can be viewed as a set of $SE(3)$-equivalent vectors in $\mathbb{R}^{3n}$ where $n = |V|$. We also consider the automorphism group of the graph, $\text{Aut}(G)$, which consists of all graph automorphisms—bijective mappings of the vertex set that preserve the adjacency relationships in the graph.

We benchmark our method against several baselines, including Torsional Diffusion (Jing et al., 2022), MCF (Wang et al., 2024), ETFlow and ETDiff (ETFlow with Gaussian prior and no special coupling) (Hassan et al., 2024) using GEOM-QM9, a subset of the GEOM dataset (Axelrod & Gomez-Bombarelli, 2022), which contains molecules with an average of 11 atoms. Following previous works, we report precision and recall for Coverage and Average Minimum RMSD (AMR) between generated and gold-standard conformers provided in CREST (Pracht et al., 2020). However, we argue that the previous threshold to classify good predictions, $\delta = 0.5$Å, is not tight enough. In fact, some important molecular properties such as energy can be very different within this range (Duan et al., 2023). Thus, to capture finer structural details, we use 0.1Å instead of 0.5Å for molecular conformer generation on GEOM-Q9. More details about these metrics are provided in Appendix C.

Table 2: Molecule conformer generation results on GEOM-QM9 ($\delta = 0.1$Å). We use the provided checkpoints and their published codes to report the performance of all baseline models.

Models	Recall				Precision				Training Time
	Coverage ↑		AMR ↓		Coverage ↑		AMR ↓		
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	
Torsional Diff.	37.7	25.0	0.178	0.147	27.6	12.5	0.221	0.195	N/A
MCF	81.9	**100.0**	0.092	0.045	78.6	**93.8**	**0.113**	0.052	N/A
ETDiff	80.7	**100.0**	0.092	0.038	71.4	75.0	0.162	0.093	42h 30m
ETDiff + Ours	81.0	**100.0**	0.089	0.037	72.2	80.0	0.151	0.077	43h 45m (↑2.9%)
ETFlow	83.8	**100.0**	0.082	0.032	77.7	90.0	0.130	0.049	63h 53m
ETFlow + Ours	**84.7**	**100.0**	**0.079**	**0.028**	**79.1**	92.5	0.122	**0.046**	64h 26m (↑0.9%)

The results for molecular conformer generation on the GEOM-QM9 dataset ($\delta = 0.1$Å) are summarized in Table 2. Our method consistently improves the performance of baseline models, as indicated by the shaded rows. Specifically, when applied to ETFlow, our approach achieves the highest mean Coverage (84.7%) and the lowest mean AMR (0.079) for Recall, as well as the best median AMR (0.028). Similarly, notable improvements are observed when augmenting ETDiff, yielding superior AMR scores compared to the baseline. While these improvements require a slight increase in training time, the overhead remains minimal, at 2.9% for ETDiff and 0.9% for ETFlow. This ensures an efficient enhancement with significant performance gains for our method.

5 CONCLUSION

In this work, we presented a novel approach to incorporating equivariance in diffusion models through a symmetrized target in the loss function. By explicitly enforcing equivariance during the optimization process, our method ensures that the learned distribution respects the desired symmetries, improving data efficiency and generalization. The use of Monte Carlo sampling to estimate the group average introduces minimal computational overhead, making our approach both scalable and efficient. Our experiments demonstrate that this method enhances sample quality compared to existing approaches. These results underscore the potential of our approach to advance the scalability and practical deployment of equivariant diffusion models across a wide range of generative tasks.

ACKNOWLEDGEMENTS

This work was funded in part by 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, and gifts from Adobe Research, Cisco Research, and Amazon. We acknowledge the support of the Stuttgart Center for Simulation Science (SimTech). VT and MN thank IMPRS-IS (International Max Planck Research School for Intelligent Systems) for the support. This work also received partial support from the Diabetes Center Berne.

REFERENCES

Emmanuel Abbe and Enric Boix-Adserà. On the non-universality of deep learning: quantifying the cost of symmetry. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6d9aac9407bcb1a5957401fa0b8de693-Abstract-Conference.html.

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*, pp. 1–3, 2024.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *CoRR*, abs/2205.15019, 2022. doi: 10.48550/ARXIV.2205.15019. URL https://doi.org/10.48550/arXiv.2205.15019.

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

Jeremiah Birrell, Markos A. Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Structure-preserving gans. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1982–2020. PMLR, 2022. URL https://proceedings.mlr.press/v162/birrell22a.html.

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. *arXiv preprint arXiv:2310.02391*, 2023.

Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale? *CoRR*, abs/2410.23179, 2024. doi: 10.48550/ARXIV.2410.23179. URL https://doi.org/10.48550/arXiv.2410.23179.

Ziyu Chen, Markos A. Katsoulakis, and Benjamin J. Zhang. Equivariant score-based generative models provably learn distributions with symmetries efficiently. *CoRR*, abs/2410.01244, 2024. doi: 10.48550/ARXIV.2410.01244. URL https://doi.org/10.48550/arXiv.2410.01244.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2990–2999. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/cohenc16.html.

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.

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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 12180–12189. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01198. URL https://doi.org/10.1109/ICCV48922.2021.01198.

Chenru Duan, Yuanqi Du, Haojun Jia, and Heather J. Kulik. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *Nat. Comput. Sci.*, 3(12): 1045–1055, 2023. doi: 10.1038/S43588-023-00563-7. URL https://doi.org/10.1038/s43588-023-00563-7.

Ahmed A. A. Elhag, T. Konstantin Rusch, Francesco Di Giovanni, and Michael M. Bronstein. Relaxed equivariance via multitask learning. *CoRR*, abs/2410.17878, 2024. doi: 10.48550/ARXIV.2410.17878. URL https://doi.org/10.48550/arXiv.2410.17878.

Fabian Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/15231a7ce4ba789d13b722cc5c955834-Abstract.html.

Mathis Gerdes, Pim de Haan, Corrado Rainone, Roberto Bondesan, and Miranda C. N. Cheng. Learning lattice quantum field theories with equivariant continuous flows. *CoRR*, abs/2207.00283, 2022. doi: 10.48550/ARXIV.2207.00283. URL https://doi.org/10.48550/arXiv.2207.00283.

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. *arXiv preprint arXiv:2303.03543*, 2023.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stark, Stephan Thaler, and Dominique Beaini. Et-flow: Equivariant flow-matching for molecular conformer generation. *arXiv preprint arXiv:2410.22388*, 2024.

Paul CD Hawkins. Conformation generation: the state of the art. *Journal of chemical information and modeling*, 57(8):1747–1756, 2017.

Lingshen He, Yuxuan Chen, Zhengyang Shen, Yiming Dong, Yisen Wang, and Zhouchen Lin. Efficient equivariant network. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 5290–5302, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/2a79ea27c279e471f4d180b08d62b00a-Abstract.html.

Jan Hermann, Zeno Schätzle, and Frank Noé. Deep neural network solution of the electronic schrödinger equation. *CoRR*, abs/1909.08423, 2019. URL http://arxiv.org/abs/1909.08423.

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.

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.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

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. *arXiv preprint arXiv:2309.17296*, 2023.

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.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2837–2845, 2021.

Vinicius Mikuni, Benjamin Nachman, and Mariel Pettee. Fast point cloud generation with diffusion models in high energy physics. *Physical Review D*, 108(3):036025, 2023.

Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.*, 22:7169–7192, 2020. doi: 10.1039/C9CP06869D. URL http://dx.doi.org/10.1039/C9CP06869D.

Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=zIUyj55nXR.

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

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, 2024a.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Rokkum Jamasb, Ilia Igashov, Weitao Du, Carla P. Gomes, Tom L. Blundell, Pietro Lio, Max Welling, Michael M. Bronstein, and Bruno E. Correia. Structure-based drug design with equivariant diffusion models. *Nat. Comput. Sci.*, 4(12):899–909, 2024b. doi: 10.1038/S43588-024-00737-X. URL https://doi.org/10.1038/s43588-024-00737-x.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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

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. *CoRR*, abs/1802.08219, 2018. URL http://arxiv.org/abs/1802.08219.

Vinh Tong, Trung-Dung Hoang, Anji Liu, Guy Van den Broeck, and Mathias Niepert. Learning to discretize denoising diffusion odes. *CoRR*, abs/2405.15506, 2024. doi: 10.48550/ARXIV.2405.15506. URL https://doi.org/10.48550/arXiv.2405.15506.

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*. OpenReview.net, 2023. URL https://openreview.net/forum?id=UaAD-Nu86WX.

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*. OpenReview.net, 2024. URL https://openreview.net/forum?id=I44Em5D5xy.

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.

Lemeng Wu, Dilin Wang, Chengyue Gong, Xingchao Liu, Yunyang Xiong, Rakesh Ranjan, Raghuraman Krishnamoorthi, Vikas Chandra, and Qiang Liu. Fast point cloud generation with straight flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9445–9454, 2023.

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*. OpenReview.net, 2022. URL https://openreview.net/forum?id=PzcvxEMzvQC.

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. *arXiv preprint arXiv:2302.02277*, 2023.

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.

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., 2022; Hermann et al., 2019), and biomolecular design (Le et al., 2022; Kozinsky et al., 2023). These networks exploit group symmetries to ensure consistent outputs under transformations like 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 learning effectively with stochastic gradient descent (SGD) (Abbe & Boix-Adserà, 2022).

Equivariance and Diffusion Models. Diffusion models have become a prominent class of generative models, particularly for tasks involving complex data distributions (Song et al., 2020a; Ho et al., 2020; Karras et al., 2022). These models transform data into noise through a forward process and learn to reverse this transformation using a neural network-based denoising function, ϕ_θ. Incorporating equivariance into diffusion models can significantly improve their performance on data with inherent symmetries, enabling more efficient learning and enhancing generalization to unseen transformations (Hoogeboom et al., 2022; Xu et al., 2022; Bose et al., 2023). To enforce equivariance, it is common to design the denoiser network ϕ_θ to be equivariant under a group G-*equivariant*. While this approach allows the learned distribution to converge to the underlying invariant distribution G-*invariant* (Chen et al., 2024), it inherits the challenges associated with equivariant network learning. Additionally, we show in this work that the Monte Carlo estimation of the loss function, following this design, suffers from high variance, which complicates the optimization process.

B Theoretical Proofs and Derivations

B.1 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 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.2 Symmetrized Diffusion Minimizer is G-equivariant

We prove that the minimizer of the symmetrized diffusion objective, ϕ_G^*, is an equivariant function of x_t. That is, $\phi_G^*(g \circ x_t, t) = g \circ \phi_G^*(x_t, t)$, for all $x_t \in \mathbb{R}^d$ and $g \in G$.

Proof. The minimizer is given by:

$$\phi_G^*(x_t, t) = \int_\Omega x_0\, q_t(x_t \mid x_0) \frac{\hat{q}^G(x_0)}{\hat{q}_t^G(x_t)}\,\mathrm{d}x_0.$$

Applying a group action $g \in G$ to x_t, we have:

$$\phi_G^*(g \circ x_t, t) = \int_\Omega x_0\, q_t(g \circ x_t \mid x_0) \frac{\hat{q}^G(x_0)}{\hat{q}_t^G(g \circ x_t)}\,\mathrm{d}x_0.$$

Next, we apply a change of variable $x_0 \mapsto g^{-1} \circ x_0$. Under this change of variable:

$$q_t(g \circ x_t \mid x_0) \to q_t(g \circ x_t \mid g \circ x_0), \quad \hat{q}^G(x_0) \to \hat{q}^G(g \circ x_0), \quad \mathrm{d}x_0 \to \mathrm{d}(g \circ x_0).$$

The integral becomes:

$$\phi_G^*(g \circ x_t, t) = \int_\Omega g \circ x_0\, q_t(g \circ x_t \mid g \circ x_0) \frac{\hat{q}^G(g \circ x_0)}{\hat{q}_t^G(g \circ x_t)}\,\mathrm{d}(g \circ x_0).$$

Using the fact that the Jacobian of the group action g is 1, the measure $\mathrm{d}(g \circ x_0)$ simplifies to $\mathrm{d}x_0$:

$$\phi_G^*(g \circ x_t, t) = \int_\Omega g \circ x_0\, q_t(g \circ x_t \mid g \circ x_0) \frac{\hat{q}^G(g \circ x_0)}{\hat{q}_t^G(g \circ x_t)}\,\mathrm{d}x_0.$$

Since the diffusion kernel $q_t(x_t \mid x_0)$ is equivariant under g, we have $q_t(g \circ x_t \mid g \circ x_0) = q_t(x_t \mid x_0)$. Similarly, the symmetrized densities satisfy $\hat{q}^G(g \circ x_0) = \hat{q}^G(x_0)$ and $\hat{q}_t^G(g \circ x_t) = \hat{q}_t^G(x_t)$. Substituting these properties, we get:

$$\phi_G^*(g \circ x_t, t) = \int_\Omega g \circ x_0\, q_t(x_t \mid x_0) \frac{\hat{q}^G(x_0)}{\hat{q}_t^G(x_t)}\,\mathrm{d}x_0.$$

Factoring out the group action g, we obtain:

$$\phi_G^*(g \circ x_t, t) = g \circ \int_\Omega x_0 \, q_t(x_t \mid x_0) \frac{\hat{q}^G(x_0)}{\hat{q}_t^G(x_t)} \, \mathrm{d}x_0.$$

Thus:

$$\phi_G^*(g \circ x_t, t) = g \circ \phi_G^*(x_t, t).$$

This completes the proof. □

B.3 Unconstrained Non-Symmetrized Diffusion Minimizer is not guaranteed G-equivaraint

On the other hand, the minimizer $\phi^*(x_t, t)$ of the following unconstrained Non-Symmetrized Diffusion loss:

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

is not guaranteed to be equivariant under the action of the symmetry group *G-equivariant*.

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

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

The gradient of the internal expectation is:

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

Setting the gradient to zero gives the minimizer:

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

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 $x_0^2 = 1$:

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

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. □

B.4 Full derivation of the group-weighted averaged posterior expectation

To compute $\phi_G^*(x_t, t)$, we start from its definition as a group-weighted posterior expectation:

$$\begin{aligned}
\phi_G^*(x_t, t) &:= \sum_{x_0 \in D} \int_G (g \circ x_0) \hat{q}_t^G(g \circ x_0 \mid x_t) \mathrm{d}\mu_G(g) \\
&= \sum_{x_0 \in D} \int_G (g \circ x_0) q_t(x_t \mid g \circ x_0) \frac{\hat{q}^G(x_0)}{\hat{q}_t^G(x_t)} \mathrm{d}\mu_G(g) \\
&= \sum_{x_0 \in D} \frac{\hat{q}^G(x_0)}{\hat{q}_t^G(x_t)} \int_G (g \circ x_0) q_t(x_t \mid g \circ x_0) \mathrm{d}\mu_G(g) \\
&= \sum_{x_0 \in D} \frac{\hat{q}^G(x_0)}{\hat{q}_t^G(x_t)} \frac{1}{(2\pi\sigma_t^2)^{d/2}} \int_G (g \circ x_0) \exp\left(-\frac{\|x_t - \alpha_t(g \circ x_0)\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu_G(g) \\
&= \sum_{x_0 \in D} \frac{1}{N} \frac{1}{\hat{q}_t^G(x_t)} \frac{1}{(2\pi\sigma_t^2)^{d/2}} \int_G (g \circ x_0) \exp\left(-\frac{\|x_t - \alpha_t(g \circ x_0)\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu_G(g) \\
&= \frac{1}{N} \frac{1}{\hat{q}_t^G(x_t)} \frac{1}{(2\pi\sigma_t^2)^{d/2}} \sum_{x_0 \in D} \int_G (g \circ x_0) \exp\left(-\frac{\|x_t - \alpha_t(g \circ x_0)\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu_G(g) \\
&= C \sum_{x_0 \in D} \int_G (g \circ x_0) \exp\left(-\frac{\|x_t - \alpha_t(g \circ x_0)\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu_G(g),
\end{aligned}$$

where $C = \frac{1}{N} \frac{1}{\hat{q}_t^G(x_t)} \frac{1}{(2\pi\sigma_t^2)^{d/2}}$ is a normalization constant independent of the integral.

We now prove that

$$C \sum_{x_0 \in D} \int_G \exp\left(-\frac{\|x_t - \alpha_t(g \circ x_0)\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu_G(g) = 1.$$

First, by the definition of C, we have

$$C \sum_{x_0 \in D} \int_G \exp\left(-\frac{\|x_t - \alpha_t(g \circ x_0)\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu_G(g) = \sum_{x_0 \in D} \int_G q_t^G(g \circ x_0 \mid x_t) \mathrm{d}\mu_G(g) = \int_\Omega \hat{q}_t^G(x_0 \mid x_t) \mathrm{d}x_0 = 1.$$

Thus, the proof is complete.

B.5 Proof of Theorem 1

Variance Reduction of the Proposed Loss

Theorem 1. *Let $\mathcal{L}_t(\phi)$ and $\bar{\mathcal{L}}_t^G(\phi)$ be the Monte Carlo estimators corresponding to Eq.* (3) *and Eq.* (4)*, respectively, where $\phi(x_t, t)$ is G-*equivariant*. Then:*

$$\mathrm{Var}(\nabla_\phi \mathcal{L}_t(\phi)) \geq \mathrm{Var}(\nabla_\phi \bar{\mathcal{L}}_t^G(\phi)), \tag{6}$$

Proof. We first state the constrained optimization loss:

$$\mathcal{L}_t(\phi) = \omega(t)\mathbb{E}_{(x_0,x_t)\sim\hat{q}_t(x_0,x_t)}\left[\|\phi(x_t,t) - x_0\|^2\right],$$
$$\text{where} \quad \phi(g \circ x_t) = g \circ \phi(x_t), \quad \forall x_t \in \Omega, \forall g \in G.$$

This constrained loss is equivalent to the symmetrized diffusion loss because:

$$\begin{aligned}
\mathcal{L}_t(\phi) &= \omega(t)\mathbb{E}_{x_t\sim\hat{q}_t(x_t)}\mathbb{E}_{x_0\sim\hat{q}_t(x_0|x_t)}\left[\|\phi(x_t,t) - x_0\|^2\right] \\
&= \omega(t)\int_\Omega \sum_{x_0\in D} \|\phi(x_t,t) - x_0\|^2 \hat{q}_t(x_0 \mid x_t)\hat{q}_t(x_t)\, dx_t \\
&= \omega(t)\int_{\Omega/G}\int_G \sum_{x_0\in D} \|\phi(g\circ x_t,t) - x_0\|^2 \hat{q}_t(x_0 \mid g\circ x_t)\hat{q}_t(g\circ x_t)\, d\mu_G(g)\, dx_t \\
&= \omega(t)\int_{\Omega/G}\int_G \sum_{x_0\in D} \|g\circ\phi(x_t,t) - x_0\|^2 \hat{q}_t(x_0 \mid g\circ x_t)\hat{q}_t(g\circ x_t)\, d\mu_G(g)\, dx_t \quad \text{(since } \phi \text{ is } G\text{-equivariant)} \\
&= \omega(t)\int_{\Omega/G}\int_G \sum_{x_0\in D} \|\phi(x_t,t) - g^{-1}\circ x_0\|^2 \hat{q}_t(x_0 \mid g\circ x_t)\hat{q}_t(g\circ x_t)\, d\mu_G(g)\, dx_t \quad \text{(since } g \text{ is isometric)} \\
&= \omega(t)\int_{\Omega/G}\int_G \sum_{x_0\in D} \|\phi(x_t,t) - g^{-1}\circ x_0\|^2 \hat{q}_t(g\circ x_t \mid x_0)\hat{q}(x_0)\, d\mu_G(g)\, dx_t \\
&= \omega(t)\int_{\Omega/G}\int_G \sum_{x_0\in D} \|\phi(x_t,t) - g^{-1}\circ x_0\|^2 \hat{q}_t(x_t \mid g^{-1}\circ x_0)\hat{q}(x_0)\, d\mu_G(g)\, dx_t \\
&= \omega(t)\int_{\Omega/G}\int_G \sum_{x_0\in D} \|\phi(x_t,t) - g^{-1}\circ x_0\|^2 \hat{q}_t(x_t \mid g^{-1}\circ x_0)\frac{1}{|D|}\, d\mu_G(g)\, dx_t \\
&= \omega(t)\int_{\Omega/G}\sum_{x_0\in D}\int_G \|\phi(x_t,t) - g^{-1}\circ x_0\|^2 \hat{q}_t(x_t \mid g^{-1}\circ x_0)\frac{1}{|D|}\, d\mu_G(g)\, dx_t \\
&= \omega(t)\int_{\Omega/G}\int_\Omega \|\phi(x_t,t) - x_0\|^2 \hat{q}_t(x_t \mid x_0)\hat{q}^G(x_0)\, dx_t \\
&= \omega(t)\mathbb{E}_{x_t\sim\hat{q}_t^G(x_t)}\mathbb{E}_{x_0\sim\hat{q}_t^G(x_0|x_t)}\left[\|\phi(x_t,t) - x_0\|^2\right].
\end{aligned}$$

Similarly, our proposed loss function, assuming $\phi(x_t,t)$ is constrained to be G-*equivariant*, is:

$$\bar{\mathcal{L}}_t^G = \mathbb{E}_{x_t\sim q_t^G(x_t)}[\|\phi(x_t,t) - \phi_G^*(x_t,t)\|^2],$$

where

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

Now, we consider the gradients of $\mathcal{L}_t$ and $\bar{\mathcal{L}}_t^G$ with respect to ϕ:

$$\nabla_\phi\mathcal{L}_t = \mathbb{E}_{(x_0,x_t)\sim\hat{q}_t^G(x_0,x_t)}2(\phi(x_t,t) - x_0),$$
$$\nabla_\phi\bar{\mathcal{L}}_t^G = \mathbb{E}_{x_t\sim q_t^G(x_t)}2(\phi(x_t,t) - \phi_G^*(x_t,t)).$$

The variances of these gradients are:

$$\mathrm{Var}(\nabla_\phi\mathcal{L}_t) = \mathrm{Var}_{(x_0,x_t)\sim\hat{q}_t^G(x_0,x_t)}[2(\phi(x_t,t) - x_0)],$$
$$\mathrm{Var}(\nabla_\phi\bar{\mathcal{L}}_t^G) = \mathrm{Var}_{x_t\sim\hat{q}_t^G(x_t)}[2(\phi(x_t,t) - \phi_G^*(x_t,t))].$$

Using the law of total variance:

$$\begin{aligned}
\mathrm{Var}(\nabla_\phi\mathcal{L}_t) &= \mathbb{E}_{x_t}\mathrm{Var}_{x_0|x_t}[2(\phi(x_t,t)) - x_0] + \mathrm{Var}_{x_t}[\mathbb{E}_{x_0|x_t}2(\phi(x_t,t) - x_0)] \\
&= \mathbb{E}_{x_t}\mathrm{Var}_{x_0|x_t}[2(\phi(x_t,t)) - x_0] + \mathrm{Var}_{x_t}[2(\phi(x_t,t) - \phi_G^*(x_t,t))] \\
&= \mathbb{E}_{x_t}\mathrm{Var}_{x_0|x_t}[2(\phi(x_t,t)) - x_0] + \mathrm{Var}(\nabla_\phi\bar{\mathcal{L}}_t^G).
\end{aligned}$$

Since variance is always non-negative, we conclude:

$$\text{Var}(\nabla_\phi \mathcal{L}_t) \geq \text{Var}(\nabla_\phi \bar{\mathcal{L}}_t^G).$$

□

C CONFORMER GENERATION 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.

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