RETHINKING DIFFUSION MODELS WITH SYMMETRIES THROUGH CANONICALIZATION WITH APPLICATIONS TO MOLECULAR GRAPH GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many generative tasks in chemistry and science involve distributions invariant to group symmetries (e.g., permutation and rotation). A common strategy enforces invariance and equivariance through architectural constraints such as equivariant denoisers and invariant priors. In this paper, we challenge this tradition through the alternative canonicalization perspective: first map each sample to an orbit representative with a canonical pose or order, train an unconstrained (non-equivariant) diffusion or flow model on the canonical slice, and finally recover the invariant distribution by sampling a random symmetry transform at generation time. Building on a formal quotient-space perspective, our work provides a comprehensive theory of canonical diffusion by proving: (i) the correctness, universality and superior expressivity of canonical generative models over invariant targets; (ii) canonicalization accelerates training by removing diffusion score complexity induced by group mixtures and reducing conditional variance in flow matching. We then show that aligned priors and optimal transport act complementarily with canonicalization and further improve training efficiency. We instantiate the framework for molecular graph generation under $S_N \times SO(3)$ symmetries. By leveraging geometric spectra-based canonicalization and mild positional encodings, canonical diffusion significantly outperforms equivariant baselines in 3D molecule generation tasks, with similar or even less computation. Moreover, with a novel architecture *Canon*, CanonFlow achieves state-of-the-art performance on the challenging GEOM-DRUG dataset, and the advantage remains large in few-step generation.

1 INTRODUCTION

Generative modeling is fundamentally grounded in the geometric structure of data. In domains such as computer vision and natural language processing (NLP), data exhibit specific symmetries—for instance, objects in images possess translation invariance, while semantic meaning in text relies on sequential order. Recent diffusion (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021) and flow-based (Liu et al., 2022; Lipman et al., 2022; Albergo et al., 2023) generative models have achieved great success across images (Dhariwal & Nichol, 2021; Rombach et al., 2022), videos (Ho et al., 2022; Singer et al., 2022), text (Li et al., 2022; Gong et al., 2022), and biomolecules (Watson et al., 2023; Corso et al., 2023). Crucially, however, modalities like images are not fully invariant to all transformations: an upside-down landscape or a reversed sentence typically loses its semantic validity. Consequently, state-of-the-art diffusion models in these fields explicitly break symmetries: they utilize fixed grid topologies or inject positional encodings (PEs) to anchor the generation process to a canonical orientation.

In contrast, molecular generation operates on a fundamentally different geometric space defined by the direct product of permutation and Euclidean symmetries, $S_N \times SE(3)$. Unlike natural images, a rotated molecule or a re-indexed molecular graph represents the exact same physical entity. Molecular generative models typically enforce the constraints by building equivariant architectures and/or invariant priors (Hoogeboom et al., 2022; Xu et al., 2022; Tian et al., 2024), so that the learned vector field (score/velocity) respects the group action and the generated distribution is symmetry-consistent. While principled, this approach often incurs substantial architectural and computational overhead (e.g., equivariant layers, tensor algebra), and it can obscure an additional challenge: symmetry

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

creates latent "gauge" ambiguity, so intermediate noisy states may correspond to multiple equivalent group-transformed configurations. This complex mixture-like nature results in "trajectory crossing" and conflicting gradients (Figure 1), as the model struggles to determine *which* valid orientation to generate from a symmetric noise vector.

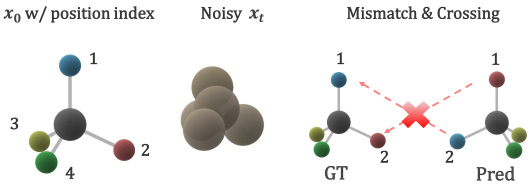


Figure 1: Motivation of CanonFlow: position-wise supervision induces mismatch loss under symmetries for equivalent molecules in diffusion models.

Motivated by this observation, we argue that the efficiency of image generation stems precisely from its lack of total invariance, and we propose to transfer this advantage to molecular modeling via **canonicalization**. In this work, we challenge the necessity of equivariant generative models and introduce a novel canonical diffusion framework, with theoretical analysis of how this symmetry breaking procedure accelerates training while ensuring validity over invariant or equivariant targets. By mapping the quotient space of molecules to a unique canonical slice, we explicitly break the symmetry, aligning the noise and data distributions. This effectively addresses the trajectory mismatch problem, transforming the complex equivariant generation task into a simplified problem on the canonical gauge.

Our starting point is the flow-matching identity that the irreducible regression error equals an expected conditional variance of endpoint displacements given an intermediate state. In Section 3, under a group-aligned lifting construction, we show this conditional variance decomposes into two nonnegative components: a within-slice term that reflects genuine transport difficulty on a canonical slice, and a symmetry-ambiguity term that arises purely from marginalizing over the latent group element. Canonicalization eliminates the symmetry-ambiguity term by fixing a gauge (working on a canonical representative per orbit), whereas optimal-transport-like couplings target the within-slice term by making the coupling more Monge-like and the trajectory closer to straight-line transport.

Based on this analysis, in Section 4 we propose a practical symmetry-aware pipeline for molecular generation that combines a canonicalizer with flexible non-equivariant backbones. We canonicalize molecules by jointly fixing permutation and rotation gauges (e.g., a Fiedler-vector rank for S_N and a rank-anchored frame for $SO(3)$), then train diffusion/flow models directly in canonical space with a gap-free prior and canonical conditions. This yields a simple yet powerful alternative to fully equivariant generative modeling: it reduces symmetry-induced ambiguity, improves few-step accuracy with the straighter learned transport, and unlocks the expressive and computational advantages of generic Transformers/GNNs for large-scale molecular generation. Furthermore, we develop a novel architecture termed ***Canon*** (Figure 7) which explicitly incorporates and refines canonical information in an additional atom hidden state, enabling canonicality-aware denoising.

Experimentally, canonicalization is compatible with any (equivariant or non-equivariant) base models. Section 5 shows that the canonicalized counterparts of baselines such as SemlaFlow (Irwin et al., 2024) yield significantly better results on popular unconditional 3D molecule generation benchmarks, including QM9 and GEOM-DRUG, with negligible computational overhead. Furthermore, leveraging the "Canon" architecture, *CanonFlow* leads to state-of-the-art molecule stability and validity metrics, outperforming the evaluated baselines by large margins. Remarkably, canonicalized models reveal faster training convergence and better few-step sampling quality.

2 Preliminaries

Canonicalization in symmetric data. We work in a measurable space $\mathcal{M}$ of structured objects. In molecular generation, we use the following state throughout the paper: $\mathbf{Z} = (\mathbf{X}, \mathbf{H}, \mathbf{A}) \in \mathcal{M}$, where $\mathbf{X} \in \mathbb{R}^{N\times 3}$ are coordinates; $\mathbf{H} \in \mathbb{R}^{N\times d_h}$ are atom features, each element consisting of d_h scalars, e.g., atom types $\{1, \dots, K\}^N$ and formal charges; $\mathbf{A} \in \mathbb{R}^{N\times N\times d_e}$ encode bonds or edge features. Let a group $\mathcal{G}$ act on $\mathcal{M}$ by

$$g \cdot \mathbf{Z} \;=\; (R,\, t,\, \pi) \cdot (\mathbf{X}, \mathbf{H}, \mathbf{A}) \;:=\; (\pi(\mathbf{X})R^\top + \mathbf{1}t^\top,\; \pi(\mathbf{H}),\; \pi(\mathbf{A})), \tag{1}$$

where $(R, t) \in SE(3)$ and $\pi \in S_N$ (simultaneous permutation of nodes and adjacency).

Definition 2.1 (Orbit and quotient)**.** The orbit of $\mathbf{Z}$ is $\mathcal{O}(\mathbf{Z}) = \{g \cdot \mathbf{Z} : g \in \mathcal{G}\}$. The quotient space is $\mathcal{M}/\mathcal{G} = \{\mathcal{O}(\mathbf{Z}) : \mathbf{Z} \in \mathcal{M}\}$.

Definition 2.2 (Invariance)**.** A probability measure μ on $\mathcal{M}$ is $\mathcal{G}$-invariant if $g\#\mu = \mu$ for all $g \in \mathcal{G}$.

We now introduce *canonicalization* as orbit selection.

Definition 2.3 (Canonicalization map and slice)**.** A measurable map $\Psi : \mathcal{M} \to \mathcal{M}$ is a *canonicalization map* if (i) $\Psi(\mathbf{Z}) \in \mathcal{O}(\mathbf{Z})$ and (ii) $\Psi(g \cdot \mathbf{Z}) = \Psi(\mathbf{Z})$ for all $g \in \mathcal{G}$. The image $S := \Psi(\mathcal{M})$ is a *canonical slice*.

Remark 2.4 (Stabilizers and non-uniqueness). If $\mathbf{Z}$ has non-trivial stabilizer $\mathrm{Stab}(\mathbf{Z}) = \{g : g \cdot \mathbf{Z} = \mathbf{Z}\}$ (e.g., graph automorphisms or symmetric geometries), then canonical representatives may be non-unique. This inherently causes discontinuities or multi-valued choices; weighted/probabilistic frames are one remedy Dym et al. (2024).

Fortunately, we assume the following holds throughout the paper, which is reasonable for (noisy) real-world 3D molecules (exact symmetries are measure-zero) and numerically verified in Table 7.

Assumption 2.5 (Free action a.s.)**.** Under p_0, the stabilizer is trivial almost surely.

There exists a finite, translation-invariant measure (Haar measure λ) that can be normalized to have total mass 1 on a compact topological group. For permutations S_N, Haar is uniform; for $SO(3)$, Haar is the uniform rotation measure; for translations, however, Haar is not finite. Thus following previous work (Hoogeboom et al., 2022; Li et al., 2025), throughout the paper we remove translations by centering (e.g., center-of-mass), leaving a compact symmetry $SO(3) \times S_N$.

Assumption 2.6 (Centered representation)**.** We assume $\mathbf{X}$ is zero-centered so global translations are removed; the remaining group is compact: $\mathcal{G} = SO(3) \times S_N$ with Haar probability λ.

Diffusion and flow-matching models. We present a continuous-time formulation, while discrete-time DDPM (Ho et al., 2020) and CTMC variants follow similarly. Let $\mathbf{Z}_0 \sim p_0$ be data. We first present score-based diffusion (Song et al., 2021). Consider a forward noising SDE on an ambient Euclidean embedding of $\mathcal{M}$ (continuous parts such as $\mathbf{X}$), a standard VP-SDE is

$$\mathrm{d}\mathbf{Z}_t = -\frac{1}{2}\beta(t)\mathbf{Z}_t\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}W_t, \tag{2}$$

with marginals p_t and score $s_\star(\mathbf{Z}, t) = \nabla_{\mathbf{Z}} \log p_t(\mathbf{Z})$ learned by a parameterized score network.

Some recent generative models are built from the equivalent flow matching viewpoint, learning a time-dependent vector field $v_t(\mathbf{Z})$ that transports a prior p_1 to data p_0 via an ODE:

$$\frac{\mathrm{d}\mathbf{Z}_t}{\mathrm{d}t} = v_t(\mathbf{Z}_t), \quad \mathbf{Z}_1 \sim p_1,\ \mathbf{Z}_0 \sim p_0. \tag{3}$$

In conditional flow matching, one specifies a coupling $\gamma(\mathbf{Z}_0, \mathbf{Z}_1)$ and a *microscopic* conditional path $\mathbf{Z}_t = \Phi_t(\mathbf{Z}_0, \mathbf{Z}_1)$ with conditional vector field $u_t(\cdot \mid \mathbf{Z}_0, \mathbf{Z}_1)$. The optimal marginal field is the conditional expectation: $v_t(\mathbf{Z}) = \mathbb{E}_\gamma\big[u_t(\mathbf{Z} \mid \mathbf{Z}_0, \mathbf{Z}_1) \mid \mathbf{Z}_t = \mathbf{Z}\big]$. The categorical parts (such as $\mathbf{A}, \mathbf{H}$), in comparison, are usually modeled by an appropriate discrete diffusion, typically with uniform/absorbing/masked states as the prior.

3 Understanding Canonicalization in Diffusion Models

Building on a formal quotient-space perspective, we provide a comprehensive theory of canonical diffusion in this section. We prove (i) a factorization theorem for invariant measures via slice distributions and Haar randomization, and the universality and superior expressivity of canonicalized generative models over invariant targets (Section 3.1); (ii) an explicit expression for the diffusion score complexity induced by group mixtures while being removable through canonicalization, and a flow-matching conditional variance decomposition showing that canonicalization reduces conditional variance, thereby accelerating diffusion/flow matching training (Section 3.2). Complete proof with mathematical notation and more discussions are available in Appendix C.

3.1 Canonical Generative Models Induce Universal Invariant Distributions

This subsection formalizes why learning on the canonical slice while sampling with Haar randomization is not only sufficient for representing any invariant target, but also more expressive with non-equivariant denoising backbones.

Universality of canonical parameterizations. First, we show that invariant distributions factor through the canonical slice. Recall that Haar probability λ is always uniform for $S_N \times SO(3)$.

Lemma 3.1 (Factorization of invariant measures; known)**.** *Suppose Assumptions 2.5 and 2.6 hold. Let μ be any $\mathcal{G}$-invariant probability measure on $\mathcal{M}$. Let Ψ be an orbit representative map defined μ-a.s., and let $\nu = \Psi\#\mu$ be the slice distribution on $S = \Psi(\mathcal{M})$. Then*

$$\mu \;=\; \int_S \left(\int_{\mathcal{G}} \delta_{g\cdot \mathbf{Z}} \, \mathrm{d}\lambda(g) \right) \mathrm{d}\nu(\mathbf{Z}). \tag{4}$$

Equivalently, if $\tilde{\mathbf{Z}} \sim \nu$, $g \sim \lambda$ independent, then $g \cdot \tilde{\mathbf{Z}} \sim \mu$.

Corollary 3.2 (Sufficiency of slice modeling)**.** *To model any invariant target μ, it suffices to model the slice distribution ν; invariance is recovered by Haar randomization.*

We propose that canonicalization is a general technique for constructing equivariant/invariant functions *without equivariant backbones*, generalizing Kaba et al. (2023):

Proposition 3.3 (Universality of canonicalized parameterizations)**.** *Let $\mathcal{G}$ act continuously and orthogonally on a compact set $K \subset \mathbb{R}^d$. Suppose we have a (measurable) canonicalization map $\Psi : K \to K$ as Definition 2.3, and a (measurable) gauge map $\kappa : K \to \mathcal{G}$ s.t.*

$$\Psi(g \cdot x) = \Psi(x),\ \kappa(g \cdot x) = g\kappa(x),\ x = \kappa(x) \cdot \Psi(x) \tag{5}$$

Consider the parametrization

$$\phi(x) = \kappa(x) \cdot f(\Psi(x)), \tag{6}$$

where f is a universal approximator on $\Psi(K)$. Then ϕ is a universal approximator of continuous $\mathcal{G}$-equivariant functions on K, and $f \circ \Psi$ is universal for continuous $\mathcal{G}$-invariant functions on K.

Stronger expressivity via symmetry breaking. We now argue that symmetry breaking with the help of non-equivariant models can be practically more expressive. Even though the true score of an invariant distribution is equivariant, enforcing equivariance inside the denoiser can **restrict architectural choices**, *limiting expressivity while increasing unnecessary computation costs* (Yan et al., 2023). Remarkably, non-equivariant models can be more expressive than their equivariant counterparts (Zhang et al., 2021; Zhou et al., 2023a). However, frame averaging or relational pooling (Puny et al., 2021; Murphy et al., 2019) are needed to recover invariant outputs, leading to a complexity of the order of the group. This motivates canonicalized diffusion: it uses a symmetry-breaking gauge (canonical order/pose) and leverages a stronger non-equivariant backbone on the slice, then restores invariance. Without the need to enumerate and average over group elements, canonicalized non-equivariant models yield stronger expressivity with similar or lower computational cost. More details are available in Appendix C.1.2.

3.2 Canonicalization Accelerates Diffusion Training

In this subsection, we provide two complementary analyses, score mixture complexity and flow-matching conditional variance, rigorously formalizing how canonicalization decreases symmetry-induced variance and accelerates diffusion model training. We further show that aligned canonical priors and optimal transport further reduce within-slice cost.

Mixture structure in diffusion score. Consider a finite group $\mathcal{G}$ with M elements acting orthogonally on $\mathbb{R}^d$ and $g_m \in \mathcal{G}$ the m-th element. Let q be a slice density supported on a canonical region S. Define the invariant mixture: $p(z_t) = \frac{1}{M}\sum_{m=1}^{M} q(g_m^{-1} \cdot z_t)$. where z_t denotes noisy states, and the summation can be substituted by integration for infinite groups. When q is differentiable and positive, the score follows given the chain rule and orthogonality,

$$\nabla \log p(z_t) = \sum_{m=1}^{M} w_m(z_t)\, g_m \cdot \nabla \log q(g_m^{-1} \cdot z_t), \quad w_m(z_t) : \;= \frac{q(g_m^{-1} \cdot z_t)}{\sum_{j=1}^{M} q(g_j^{-1} \cdot z_t)}. \tag{7}$$

Where multiple group copies overlap, the responsibilities $w_m(z_t)$ vary rapidly. The induced sharp score fields require more capacity and smaller step sizes for stable reverse integration. As noted by Yan et al. (2023), invariant training leads to increased modes and mixture-like optimal denoising scores. Instead, canonicalization removes the mixture: there is a single component on the slice ($M = 1$), which smooths the score landscape, simplifying denoising score network training.

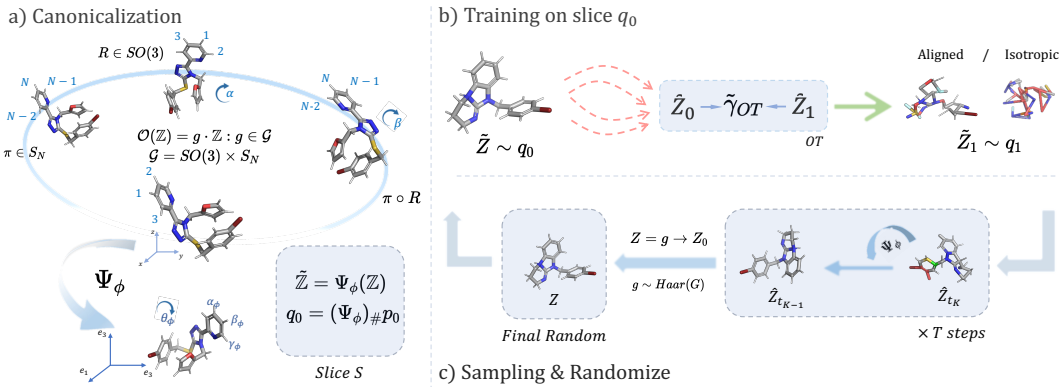


Figure 2: Overview of our canonicalized generation pipeline. (a) Canonicalization: map a molecule Z to a slice representative $\tilde{Z} = \Psi_\phi(Z)$ under $\mathcal{G} = SO(3) \times S_N$, inducing $q_0 = (\Psi_\phi)_\# p_0$. (b) Training: learn a diffusion/flow model on $\tilde{Z}_0 \sim q_0$ with slice prior q_1 (optionally using an OT coupling for aligned pairings). (c) Sampling: generate on the slice (optionally with projected canonical sampling) and then apply $g \sim \mathrm{Haar}(\mathcal{G})$ to obtain an invariant sample $Z = g \cdot \hat{Z}_0$.

Conditional flow variance reduction via symmetry ambiguity elimination. Here we formalize our central result: *canonicalization reduces conditional variance in flow matching training* by eliminating symmetry ambiguity. Without loss of generality, consider the linear-path setting, the conditional vector field is constant:

$$\mathbf{Z}_t = (1-t)\mathbf{Z}_0 + t\mathbf{Z}_1, \quad (\mathbf{Z}_0, \mathbf{Z}_1) \sim \gamma, \quad u_t(\mathbf{Z}_t \mid \mathbf{Z}_0, \mathbf{Z}_1) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Z}_t = \mathbf{Z}_1 - \mathbf{Z}_0. \tag{8}$$

The optimal marginal field is $v_t(\mathbf{Z}) = \mathbb{E}[\mathbf{Z}_1 - \mathbf{Z}_0 \mid \mathbf{Z}_t = \mathbf{Z}]$. Let $\hat{v}$ be any measurable predictor of $v_t(\mathbf{Z}_t)$ from $(t, \mathbf{Z}_t)$. Then the minimum achievable MSE (equivalently, the *irreducible flow-matching error*) is the conditional variance: $\inf_{\hat{v}} \mathbb{E}\big[\, \|\hat{v}(t, \mathbf{Z}_t) - (\mathbf{Z}_1 - \mathbf{Z}_0)\|^2 \,\big] = \mathbb{E}\big[\mathrm{Var}(\mathbf{Z}_1 - \mathbf{Z}_0 \mid t, \mathbf{Z}_t)\big]$.

We aim to understand when $\mathrm{Var}(Z_1 - Z_0 | t, Z_t)$ is large, especially in the presence of symmetry. Consider the following *symmetry-mixture data model*. Let a compact group $\mathcal{G}$ act linearly and orthogonally on $\mathbb{R}^d$ (so $\|g \cdot x\| = \|x\|$). Assume the data distribution p_0 is $\mathcal{G}$-invariant:

$$Z_0 \overset{d}{=} G \cdot \widetilde{Z}_0 \sim p_0, \; G \sim \lambda \text{ (Haar)}, \; \widetilde{Z}_0 \sim q_0, \; \widetilde{Z}_0 = \Psi(Z_0), \; q_0 = \Psi_\# p_0, \tag{9}$$

where Ψ is a (measurable) canonicalizer mapping each orbit to a representative on a slice. Intuitively, without canonicalization the posterior $p(Z_0 \mid Z_t)$ can be multi-modal over the latent symmetry element G, inflating conditional variance.

Compare the following two training paradigms: (i) **canonical slice training** (quotient space), train on canonicalized data $\widetilde{Z}_0 := \Psi(Z_0) \sim q_0$ with a slice prior $\widetilde{Z}_1 \sim q_1$ and a coupling $\widetilde{\gamma}(\widetilde{Z}_0, \widetilde{Z}_1)$; (ii) **invariant-mixture training** (ambient space), train directly on the invariant data $Z_0 \sim p_0$ with an ambient prior $Z_1 \sim p_1$ and a coupling $\gamma(Z_0, Z_1)$. To relate slice coupling to invariant ambient-space coupling, we define the *group-aligned lift* of a given $\widetilde{\gamma}$:

$$G \sim \lambda, \; (\widetilde{Z}_0, \widetilde{Z}_1) \sim \widetilde{\gamma} \text{ independent}, \quad (Z_0, Z_1) := (G \cdot \widetilde{Z}_0, \; G \cdot \widetilde{Z}_1) \sim \gamma^{\mathrm{lift}}. \tag{10}$$

We emphasize that the coupling correspondence (namely the invariant training uses the induced group-aligned lifting as the ambient coupling) is practically reasonable. When $\tilde{Z}_1 \sim \mathcal{N}(0, I)$ and the slice coupling is the independent product $\tilde{\gamma} := q_0 \otimes q_1$, then the ambient prior $Z_1 \sim p_1 = \mathcal{N}(0, I)$ and the invariant coupling recovers the widely adopted product coupling (w/o optimal transport) $\gamma_{mix} := p_0 \otimes p_1$; refer to Proposition C.12 for more details. This generally does not hold, however, if $\tilde{Z}_1$ is group-aware, or $\tilde{Z}_1$ and $\tilde{Z}_0$ are not independent; more details are available in Appendix C.2.2.

Remark 3.4. All variance decompositions below compare *under the lifted coupling* γ^{lift} (i.e., they isolate the effect of marginalizing over G versus conditioning on G).

We have the following sharp conditional variance comparison under the lifted coupling.

Theorem 3.5 (Variance decomposition under group-aligned lift)**.** *Assume $\mathcal{G}$ acts orthogonally. Under the group-aligned lifted coupling $G \sim \lambda$ independent of $(\widetilde{Z}_0, \widetilde{Z}_1)$ and $(Z_0, Z_1) = (G \cdot \widetilde{Z}_0, G \cdot \widetilde{Z}_1)$, let $Z_t = (1-t)Z_0 + tZ_1$ and $\widetilde{Z}_t = (1-t)\widetilde{Z}_0 + t\widetilde{Z}_1$. With $U := Z_1 - Z_0 = G \cdot \Delta$ and $\Delta := \widetilde{Z}_1 - \widetilde{Z}_0$,*

$$\mathrm{Var}(U \mid Z_t) = \underbrace{\mathbb{E}\Big[\mathrm{Var}(\Delta \mid \widetilde{Z}_t) \mid Z_t\Big]}_{\textit{within-slice difficulty}} + \underbrace{\mathrm{Var}(\mathbb{E}[U \mid Z_t, G] \mid Z_t)}_{\textit{symmetry ambiguity} \geq 0}. \tag{11}$$

Consequently, invariant training without canonicalization admits larger conditional variance:

$$\mathbb{E}\big[\mathrm{Var}(U \mid Z_t)\big] \;\geq\; \mathbb{E}\big[\mathrm{Var}(\Delta \mid \widetilde{Z}_t)\big]. \tag{12}$$

In particular, canonicalization eliminates the second term caused by symmetry ambiguity, and strict inequality typically holds whenever G remains ambiguous given Z_t and the conditional mean drift depends on G; see concrete posterior collision bounds in Appendix C.2.2. The decomposition is empirically validated on practical datasets in Table 9.

Remark 3.6 (Only non-equivariant models benefit from canonicalization). Notably, Theorem 3.5 only provides the lower bound on the irreducible error, but does not indicate that all models are able to achieve this lower bound in practice. We emphasize that only non-equivariant models benefit from symmetry ambiguity elimination practically since canonicalization and canonical conditions expand the effective hypothesis class, echoing our analysis in Section 3.1. Equivariant models, by contrast, retain the same ambient Bayes risk after canonicalization as in invariant training. Further details are given in Proposition C.15 in the Appendix.

Within-slice simplification with aligned canonical prior and Monge coupling. Theorem 3.5 isolates two distinct sources of irreducible error: within-slice difficulty and symmetry ambiguity. Canonicalization reduces the second term, but the first term still depends on the slice coupling $\tilde{\gamma}(\tilde{Z}_0, \tilde{Z}_1)$ and can remain large when the slice coupling is far from straight transport. We hereby present techniques targeting the first term: *aligned canonical priors* and *optimal transport*.

We start from the principles of aligned prior $\tilde{Z}_1 \sim q_1$ for canonical slice. Consider the case where one keeps a *misaligned* simple prior (e.g. $q_1 = \mathcal{N}(0, I)$ while q_0 on the canonical slice is far from centered/isotropic or is low-dimensional), then the slice coupling is still difficult and the first term in Theorem 3.5 can remain large. Therefore, we propose two simple yet effective approaches to derive **aligned canonical priors**: (i) *learnable priors* within certain function classes (e.g. a Gaussian with learnable mean and variance); (ii) *KL projections of* q_0 onto certain function classes (e.g. moment-matched Gaussian). For example, among Gaussian canonical priors, the moment-matched Gaussian is $q_1^\star = \mathcal{N}(\mathbb{E}_{q_0}[\widetilde{Z}_0], \mathrm{Cov}_{q_0}(\widetilde{Z}_0))$ (Proposition C.20).

Given the canonical data q_0 and the prior q_1, the slice conditional variance $\mathbb{E}[\mathrm{Var}(\widetilde{Z}_1 - \widetilde{Z}_0 \mid \widetilde{Z}_t)]$ still depends on the coupling $\tilde{\gamma}$. In favorable regimes, e.g. Monge-like optimal transport (OT) under standard regularity conditions, this term can be small. Remarkably, OT has been applied in some previous work (Irwin et al., 2024), and we now formally clarify its effects on symmetry space. While OT generally reduces trajectory intersections and conditional variances regardless of whether canonicalization is applied, the solutions are not unique or stable in the presence of symmetry. Canonicalization stabilizes the solution by providing a gauge, and further benefits non-equivariant models with canonical conditions, as illustrated in Appendix C.2.2. Alternatively, canonicalization serves as an "implicit OT" by aligning the data-noise pair into the same gauge.

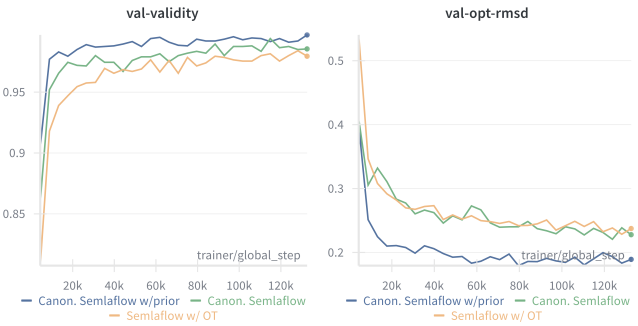


Figure 3: Training trajectories of our canonicalized model (visualization of the learning dynamics).

Remark 3.7 (Canonicalization and optimal transport act complementarily). In the presence of symmetry, canonicalization eliminates group ambiguity, while optimal transport reduces within-slice difficulty. OT solutions can be closer to unique/Monge-like with canonicalization (Figure 5).

We finally clarify conditions for training-sampling consistency. Suppose the network is conditioned on an auxiliary variable C (e.g. "canonical rank" positional encoding, a frame ID, or any deterministic or stochastic output of a canonicalizer). Training then implicitly targets a conditional field $\widetilde{v}(t, z, c)$ under a joint law $\tilde{\pi}_t(c, z)$ induced by the training pipeline. We say there is *no condition mismatch at the start* if the inference-time joint law $\tilde{\pi}_1^{\mathrm{inf}}(C, \widetilde{Z}_1)$ equals the training-time joint law $\tilde{\pi}_1^{\mathrm{tr}}(C, \widetilde{Z}_1)$. With either (i) aligned canonical priors, or (ii) isotropic priors and training-time OT, one keeps the training–sampling consistency. The training trajectories in Figure 3 strongly support our claim that canonicalization, especially when combined with aligned priors, significantly accelerates training.

4 Canonical Diffusion for Molecular Graph Generation

Building on the theoretical foundations established in Section 3, we now instantiate our canonical molecular graph diffusion framework under the joint symmetry group $\mathcal{G} = S_N \times SO(3)$ (atom index permutations and global rotations).

4.1 Canonicalization of Molecular Graphs

The key design choice lies in constructing a canonicalizer that is both *geometrically principled* and *computationally tractable*. We decompose the full symmetry group, addressing permutation and rotation symmetries sequentially:

$$\Psi_\phi(\mathbf{Z}) \;=\; \Psi_\phi^{(\mathrm{rot})}\Big(\Psi_\phi^{(\mathrm{perm})}(\mathbf{Z})\Big), \tag{13}$$

where $\Psi_\phi^{(\mathrm{perm})}$ establishes a canonical atom ordering and $\Psi_\phi^{(\mathrm{rot})}$ fixes a global reference frame.

We first consider **canonicalization via geometric spectra**, which breaks permutation symmetry of atom indices in a principled manner. The Fiedler vector, i.e. the eigenvector corresponding to the second smallest eigenvalue (the algebraic connectivity) of a graph's Laplacian matrix, determines a canonical atom ordering in a natural way. Concretely, given a molecule, we build a symmetric kernel matrix $\mathbf{W}(\mathbf{X}, \mathbf{A}) \in \mathbb{R}^{N\times N}$ from pairwise distances (e.g., $W_{ij} = k(\|\mathbf{X}_i - \mathbf{X}_j\|)$ for a radial kernel k) to handle the 3D geometry, forming the normalized geometric Laplacian

$$\mathbf{L}(\mathbf{X}, \mathbf{A}) \;=\; \mathbf{D}^{-1}(\mathbf{D} - \mathbf{W}), \quad \mathbf{D} = \mathrm{diag}(\mathbf{W}\mathbf{1}). \tag{14}$$

$\mathbf{L}$ depends only on distances and is rotation-invariant. We then use the Fiedler vector $\mathbf{u_2} \in \mathbb{R}^N$ to define a canonical ordering

$$\pi^\star(\mathbf{Z}) \;:=\; \mathrm{argsort}(\mathbf{u_2}) \in S_N. \tag{15}$$

Under the standard mass–spring interpretation of the Laplacian, this corresponds to the lowest-frequency non-rigid vibration mode Knyazev (2018). This rank yields the classical spectral bisection and provides a principled seriation keeping strongly connected substructures contiguous Pothen et al. (1990). In the molecular context, the spectral ordering serves as a *geometry-aware linearization*.

Finally, we optionally define a global $SO(3)$ gauge for global rotations in a rank-consistent way after applying $\pi^\star$. We choose a small set of anchor atoms in the canonical order (e.g., the two extremes define a primary axis and a third anchor determines a plane normal), and rotate the centered coordinates into the resulting right-handed orthonormal frame. This yields a deterministic representative under the joint $S_N \times SO(3)$ action; the full algorithm is deferred to Algorithm 2.

There are also other existing canonicalization methods, mostly for permutation in abstract graphs (Zhao et al., 2024; Ma et al., 2023a; Dong et al., 2024) instead of geometric graphs, while existing OT methods mostly target coordinates; our method consistently considers the joint group in a stable manner. Different canonicalization methods may differ in stability and practical performance, yet we experimentally find that our novel geometric spectra canonicalization outperforms others (Appendix D.1).

4.2 Canonical Diffusion and CanonFlow

Canonicality-aware denoising architectures. We elaborate on two types of architectural changes to enable non-equivariant denoising models in canonical diffusion: (i) *positional encoding only*,

where we keep the original model architectures except that the model is conditioned on canonical rank embeddings $C = \varphi(r_i)$ with normalized rank $r_i = (i-1)/N$ to handle various molecular sizes; (ii) ***Canon*** *architecture*, our newly designed architecture that explicitly keeps atom canonical embeddings as a learnable hidden state $\mathbf{C} \in \mathbb{R}^{N \times d_c}$, which interacts with the embeddings of $\mathbf{X}, \mathbf{H}, \mathbf{A}$ in each layer (Figure 7); see more details in Appendix D.2. We use the prefix "*Canon.*" to indicate the first type, where everything is consistent with the base model except the PE; "*CanonFlow*" specifically refers to the flow matching model with *Canon* architecture.

Training on the canonical slice. Training proceeds via standard diffusion or flow matching procedures on the canonical slice $\widetilde{\mathbf{Z}} = \Psi_\phi(\mathbf{Z})$ and the slice distribution $q_0 = (\Psi_\phi)\#p_0$. For convenience, we illustrate with the flow matching framework, and diffusion models follow analogously. Given a coupling $\widetilde{\gamma}(\widetilde{\mathbf{Z}}_0, \widetilde{\mathbf{Z}}_1)$ between canonicalized data $\widetilde{\mathbf{Z}}_0 \sim q_0$ and the (optionally aligned) prior $\widetilde{\mathbf{Z}}_1 \sim q_1$, we construct interpolated paths $\widetilde{\mathbf{Z}}_t = \Phi_t(\widetilde{\mathbf{Z}}_0, \widetilde{\mathbf{Z}}_1)$ for $t \sim \mathrm{Unif}[0,1]$. The model m_θ learns to predict the conditional velocity field: $\min_\theta \ \mathbb{E}\Big[\big\|m_\theta(\widetilde{\mathbf{Z}}_t, t;\, C) - u_t(\widetilde{\mathbf{Z}}_t \mid \widetilde{\mathbf{Z}}_0, \widetilde{\mathbf{Z}}_1)\big\|^2\Big]$, where $u_t = \frac{\mathrm{d}}{\mathrm{d}t}\Phi_t$ including velocities for both continuous and discrete features. As explained in Section 3.2, we also optionally adopt an aligned prior, or enable OT coupling early in training and anneal its usage over epochs (termed *OT anneal*) to stabilize optimization.

Sampling and restoring invariance. Starting from the slice prior $\widetilde{\mathbf{Z}}_1 \sim q_1$ (the same marginal prior used in training), we integrate the learned dynamics to obtain a canonical sample $\widehat{\mathbf{Z}}_0$ on the slice S; no OT matching is required at inference. Denoting the sampled canonical distribution by $\tilde{\mu}$, we can then restore the invariance by sampling $g \sim \lambda$, $\widetilde{\mathbf{Z}} \sim \tilde{\mu}$ and outputting $g \cdot \widetilde{\mathbf{Z}}$, thus the randomized distribution $\mu := \int_{\mathcal{G}} (g\cdot)\#\tilde{\mu}\, \mathrm{d}\lambda(g)$ is invariant (Proposition D.1).

Improved techniques with canonicality. A subtle discrepancy arises between training and inference: canonical ranks are computed exactly from $\Psi_\phi(\mathbf{Z})$ during training, whereas at inference the model generates from noise without access to ground-truth ranks. Moreover, following Dym et al. (2024), continuous canonicalization is impossible for S_N and $SO(d)$. To address this issue and to mitigate the train-test gap, we relax the ranks into continuous values and inject small perturbations to rank features during training, with an optional auxiliary head that self-estimates ranks from intermediate representations. In sampling, models either utilize fixed canonical conditions, or adaptively re-estimate the ranks to align the noisy data and the canonical conditions, which we term *projected canonical sampling* (PCS). Inspired by Li et al. (2025), we also randomly drop canonical conditions with probability p_{drop} during training, and enable classifier-free guidance (CFG) (Ho & Salimans, 2022) extrapolating conditional and unconditional generation during inference. More training and sampling techniques are deferred to Appendix D.

5 Experiments

Experimental Setup. We evaluate our canonicalization approach on two widely adopted benchmarks for unconditional 3D molecule generation: QM9 (Ramakrishnan et al., 2014) and GEOM-DRUG (Axelrod & Gomez-Bombarelli, 2022). While QM9 consists of relatively small molecules, GEOM-DRUG provides a more challenging testbed that better differentiates the capabilities of various generative models. We adopt identical train/validation/test splits as prior work (Vignac et al., 2023; Le et al., 2023). Following Irwin et al. (2024), we exclude molecules containing more than 72 atoms from the GEOM-DRUG training set for computational efficiency, which accounts for approximately 1% of the data; the validation and test sets remain intact. During evaluation, we sample molecule sizes from the empirical distribution in the test set and generate the corresponding number of atoms by numerically integrating the learned flow dynamics. All reported metrics are computed over 10k generated molecules with 3 independent runs. Unless otherwise specified, we adopt the settings and hyper-parameters in SemlaFlow (Irwin et al., 2024), which serves as our base model. By default, canonicalization is conducted on the S_N group; results of canonicalizing $S_N \times SO(3)$ as well as *more ablation studies* are deferred to Appendix E.

Main results. Applying the canonical diffusion framework to SemlaFlow and CanonFlow, we achieve state-of-the-art results on most metrics across both datasets.

Table 1: Performance of canonicalized SemlaFlow on the QM9 dataset. Atom stability, molecule stability, and validity are reported as percentages; Opt-RMSD is reported in Å.

Methods	Atom Stab ↑	Mol Stab ↑	Valid ↑	Opt-RMSD ↓	NFE ↓
EDM	98.7	82.0	91.9	–	1000
GCDM	98.7	85.7	94.8	–	1000
MUDiff	98.3	89.9	95.3	–	1000
FlowMol	99.7	96.2	97.3	–	100
MiDi	99.8	97.5	97.9	–	500
EQGAT-diff	$99.9_{\pm 0.0}$	$98.7_{\pm 0.18}$	$99.0_{\pm 0.16}$	–	500
SemlaFlow w/ OT	$99.9_{\pm 0.0}$	$99.56_{\pm 0.07}$	$99.31_{\pm 0.08}$	$0.23_{\pm 0.00}$	100
Canon. SemlaFlow w/ OT anneal + PCS (ours)	$99.9_{\pm 0.0}$	$\underline{99.62}_{\pm 0.08}$	$\underline{99.49}_{\pm 0.06}$	$\underline{0.21}_{\pm 0.00}$	100
Canon. SemlaFlow w/ Prior + PCS (ours)	$99.9_{\pm 0.0}$	$\mathbf{99.64}_{\pm 0.04}$	$\mathbf{99.58}_{\pm 0.06}$	$\mathbf{0.17}_{\pm 0.00}$	100

Table 2: Performance of canonicalized SemlaFlow and CanonFlow on the GEOM-DRUG dataset. Atom stability, molecule stability, validity, as well as uniqueness and novelty are reported as percentages.

Methods	Atom Stab ↑	Mol Stab ↑	Valid ↑	Unique ↑	Novel ↑	NFE ↓
FlowMol	99.0	67.5	51.2	-	-	100
MiDi	99.8	91.6	77.8	100.0	100.0	500
EQGAT-diff	$99.8_{\pm 0.0}$	$93.4_{\pm 0.21}$	$94.6_{\pm 0.24}$	$100.0_{\pm 0.0}$	$99.9_{\pm 0.07}$	500
SemlaFlow w/ OT	$99.8_{\pm 0.0}$	$97.3_{\pm 0.08}$	$93.9_{\pm 0.19}$	$100.0_{\pm 0.0}$	$99.6_{\pm 0.03}$	100
Canon. SemlaFlow w/ OT anneal (ours)	$99.8_{\pm 0.0}$	$\underline{98.1}_{\pm 0.03}$	$\underline{95.0}_{\pm 0.20}$	$100.0_{\pm 0.0}$	$99.7_{\pm 0.02}$	100
CanonFlow w/ OT anneal (ours)	$\mathbf{99.9}_{\pm 0.0}$	$\mathbf{98.4}_{\pm 0.02}$	$\mathbf{95.9}_{\pm 0.08}$	$100.0_{\pm 0.0}$	$99.7_{\pm 0.01}$	100

Table 1 shows results on QM9. While all methods achieve near-saturated atom stability, our canonicalized variants improve upon SemlaFlow in both molecule stability and validity. Notably, Canon. SemlaFlow (Prior + PCS) achieves the lowest Opt-RMSD (a 26% reduction compared to the baseline), indicating that canonicalization guides the model toward energy-minimized conformations. The training dynamics on QM9 illustrated in Figure 3 also demonstrate that canonicalization significantly accelerates training convergence.

Table 2 presents results on the more challenging GEOM-DRUG dataset, which contains larger and structurally diverse drug-like molecules that better differentiate model capabilities. On this benchmark, our canonicalized approach demonstrates clear improvements over baselines. In particular, Canon. SemlaFlow consistently improves validity and molecule stability over the SemlaFlow baseline. Remarkably, our CanonFlow achieves **SOTA** performance across atom stability, molecule stability and validity – the most important metrics, outperforming all previous models by a large margin (exceeding the SemlaFlow baseline by 1.1 percentage points in molecule stability and 2.0 percentage points in validity).

We also report *few-step generation* performance in Appendix E.2. The advantages of canonicalized models are still significant with only 50 steps, one tenth of the 500 steps used by several baselines. Notably, our method introduces almost no computational or sampling-time overhead (Appendix E.1.3), validating the efficient and effective guidance signal of canonical conditioning in sampling.

6 Conclusion

We studied symmetry in diffusion and flow-based generative modeling through canonicalization. Under group-aligned lifting, symmetry induces latent *group ambiguity* at intermediate noise levels, increasing conditional variance. Canonicalization removes this ambiguity term and enables expressive non-equivariant backbones. Aligned priors and near-Monge couplings can mitigate the remaining within-slice difficulty. Canonical diffusion yields state-of-the-art results across several metrics on popular geometric molecule generation benchmarks, with significant advantages in training convergence. While the current experimental scope of this paper lies in unconditional small molecule generation, we hope to extend this framework to other biomolecular (e.g., protein) structure prediction and conditional generation (e.g., binder design) tasks with more potential symmetry groups in the future.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, manuscript-format checks. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper presents work whose goal is to advance machine learning, generative modeling, and molecular design. There are some potential societal consequences of our work, none of which we feel must be specifically highlighted here.

REPRODUCIBILITY STATEMENT

We provide all assumptions and complete proofs for our theoretical results in Appendix C. All details to reproduce our experiments, including datasets, algorithms, architectures, and experimental settings, are included in Appendix D and Appendix E.

REFERENCES

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

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

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

Deyu Bo, Chuan Shi, Lele Wang, and Renjie Liao. Specformer: Spectral graph neural networks meet transformers. *arXiv preprint arXiv:2303.01028*, 2023.

Chen Cai, Truong Son Hy, Rose Yu, and Yusu Wang. On the connection between MPNN and graph transformer. In *International conference on machine learning*, pp. 3408–3430. PMLR, 2023.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=DmT862YAieY.

Nicola De Cao and Thomas Kipf. MolGAN: An implicit generative model for small molecular graphs, 2022. URL https://arxiv.org/abs/1805.11973.

Xiaohui Chen, Jiaxing He, Xu Han, and Li-Ping Liu. Efficient and degree-guided graph generation via discrete diffusion modeling. *arXiv preprint arXiv:2305.04111*, 2023.

Gabriele Corso, Hannes St"ark, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations*, 2023.

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

Yuhui Ding and Thomas Hofmann. Scalable non-equivariant 3D molecule generation via rotational alignment. *arXiv preprint arXiv:2506.10186*, 2025.

Zehao Dong, Muhan Zhang, Philip Payne, Michael Province, Carlos Cruchaga, Tianyu Zhao, Fuhai Li, and Yixin Chen. Rethinking the power of graph canonization in graph representation learning with stability. In *International Conference on Learning Representations*, 2024.

Ian Dunn and David Ryan Koes. Mixed continuous and categorical flow matching for 3D de novo molecule generation. *arXiv preprint arXiv:2404.19739*, 2024. URL https://arxiv.org/abs/2404.19739.

Alexandre Agm Duval, Victor Schmidt, Alex Hernández-Garcıa, Santiago Miret, Fragkiskos D Malliaros, Yoshua Bengio, and David Rolnick. FAENet: Frame averaging equivariant GNN for materials modeling. In *International Conference on Machine Learning*, pp. 9013–9033. PMLR, 2023.

Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations. *arXiv preprint arXiv:2110.07875*, 2021.

Nadav Dym, Hannah Lawrence, and Jonathan W Siegel. Equivariant frames and the impossibility of continuous canonicalization. *arXiv preprint arXiv:2402.16077*, 2024.

Wei Feng, Lvwei Wang, Zaiyun Lin, Yanhao Zhu, Han Wang, Jianqiang Dong, Rong Bai, Huting Wang, Jielong Zhou, Wei Peng, Bo Huang, and Wenbiao Zhou. Generation of 3D molecules in pockets via language model, 2023. URL https://arxiv.org/abs/2305.10133.

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

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. DiffuSeq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stark, Stephan Thaler, and Dominique Beaini. ET-Flow: Equivariant flow-matching for molecular conformer generation, 2024. URL https://arxiv.org/abs/2410.22388.

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

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Haokai Hong, Wanyu Lin, and Kay Chen Tan. Accelerating 3D molecule generation via jointly geometric optimal transport, 2025. URL https://arxiv.org/abs/2405.15252.

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.

Xiaoyang Hou, Tian Zhu, Milong Ren, Dongbo Bu, Xin Gao, Chunming Zhang, and Shiwei Sun. Improving molecular graph generation with flow matching and optimal transport, 2024. URL https://arxiv.org/abs/2411.05676.

Chenqing Hua, Sitao Luan, Minkai Xu, Rex Ying, Jie Fu, Stefano Ermon, and Doina Precup. MUDiff: Unified diffusion for complete molecule generation, 2024. URL https://arxiv.org/abs/2304.14621.

Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2D & 3D diffusion models for complete molecule generation, 2023. URL https://arxiv.org/abs/2305.12347.

Yinan Huang, Xingang Peng, Jianzhu Ma, and Muhan Zhang. Boosting the Cycle Counting Power of Graph Neural Networks with I^2-GNNs. *arXiv e-prints*, art. arXiv:2210.13978, October 2022. doi: 10.48550/arXiv.2210.13978.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. SemlaFlow–efficient 3D molecular generation with latent attention and equivariant flow matching. *arXiv preprint arXiv:2406.07266*, 2024.

Yunhui Jang, Dongwoo Kim, and Sungsoo Ahn. Hierarchical graph generation with k^2-trees. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Keyue Jiang, Jiahao Cui, Xiaowen Dong, and Laura Toni. Bures-wasserstein flow matching for graph generation, 2025. URL https://arxiv.org/abs/2506.14020.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation, 2019. URL https://arxiv.org/abs/1802.04364.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15546–15566. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kaba23a.html.

Jinwoo Kim, Tien Dat Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. Pure transformers are powerful graph learners. *arXiv preprint arXiv:2207.02505*, 2022.

Andrew V. Knyazev. On spectral partitioning of signed graphs, 2018. URL https://arxiv.org/abs/1701.01394.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step, 2024. URL https://arxiv.org/abs/2403.13117.

Hannah Lawrence, Elyssa Hofgard, Yuxuan Chen, Tess Smidt, and Robin Walters. Detecting symmetry-breaking in molecular data distributions. In *AI for Accelerated Materials Design-ICLR 2025*, 2025a.

Hannah Lawrence, Vasco Portilheiro, Yan Zhang, and Sékou-Oumar Kaba. Improving equivariant networks with probabilistic symmetry breaking. *arXiv preprint arXiv:2503.21985*, 2025b.

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.

Joongwon Lee, Seonghwan Kim, Seokhyun Moon, Hyunwoo Kim, and Woo Youn Kim. FragFM: Hierarchical framework for efficient molecule generation via fragment-level discrete flow matching, 2025. URL https://arxiv.org/abs/2502.15805.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-LM improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Zian Li, Xiyuan Wang, Yinan Huang, and Muhan Zhang. Is distance matrix enough for geometric deep learning? In *Advances in Neural Information Processing Systems*, volume 36, pp. 37413–37447, 2023.

Zian Li, Xiyuan Wang, Shijia Kang, and Muhan Zhang. On the completeness of invariant geometric deep learning models. *arXiv preprint arXiv:2402.04836*, 2024.

Zian Li, Cai Zhou, Xiyuan Wang, Xingang Peng, and Muhan Zhang. Geometric representation condition improves equivariant molecule generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36921–36953. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25dz.html.

Derek Lim, Joshua Robinson, Lingxiao Zhao, Tess Smidt, Suvrit Sra, Haggai Maron, and Stefanie Jegelka. Sign and basis invariant networks for spectral graph representation learning. *arXiv preprint arXiv:2202.13013*, 2022.

Yuchao Lin, Jacob Helwig, Shurui Gui, and Shuiwang Ji. Equivariance via minimal frame averaging for more symmetries and efficiency. *arXiv preprint arXiv:2406.07598*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Peter Lippmann, Gerrit Gerhartz, Roman Remme, and Fred A Hamprecht. Beyond canonicalization: How tensorial messages improve equivariant message passing. *arXiv preprint arXiv:2405.15389*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Tianze Luo, Zhanfeng Mo, and Sinno Jialin Pan. Fast graph generation via spectral diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Youzhi Luo and Shuiwang Ji. An autoregressive flow model for 3D molecular geometry generation from scratch. In *International conference on learning representations (ICLR)*, 2022.

Youzhi Luo, Keqiang Yan, and Shuiwang Ji. GraphDF: A discrete flow model for molecular graph generation, 2021. URL https://arxiv.org/abs/2102.01189.

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. *Advances in Neural Information Processing Systems*, 37: 60936–60979, 2024.

Jiangyan Ma, Yifei Wang, and Yisen Wang. Laplacian canonization: A minimalist approach to sign and basis invariant spectral embedding. *arXiv preprint arXiv:2310.18716*, 2023a.

Liheng Ma, Chen Lin, Derek Lim, Adriana Romero-Soriano, Puneet K. Dokania, Mark Coates, Philip Torr, and Ser-Nam Lim. Graph inductive biases in transformers without message passing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23321–23337. PMLR, 2023b. URL https://proceedings.mlr.press/v202/ma23c.html.

Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably powerful graph networks. *arXiv preprint arXiv:1905.11136*, 2019.

Alex Morehead and Jianlin Cheng. Geometry-complete diffusion for 3D molecule generation and optimization. *Communications Chemistry*, 7(1):150, 2024.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 4602–4609, 2019.

Christopher Morris, Gaurav Rattan, and Petra Mutzel. Weisfeiler and Leman go sparse: Towards scalable higher-order graph embeddings. *Advances in Neural Information Processing Systems*, 33: 21824–21840, 2020.

Ryan Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Relational pooling for graph representations. In *International Conference on Machine Learning*, pp. 4663–4673. PMLR, 2019.

Alex Pothen, Horst D. Simon, and Kang-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11(3):430–452, 1990. doi: 10.1137/0611030.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *arXiv preprint arXiv:2110.03336*, 2021.

Yiming Qin, Manuel Madeira, Dorina Thanou, and Pascal Frossard. DeFoG: Discrete flow matching for graph generation. *arXiv preprint arXiv:2410.04263*, 2024.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):140022, 2014.

Ladislav Rampasek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and D. Beaini. Recipe for a general, powerful, scalable graph transformer. *arXiv preprint arXiv:2205.12454*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. GraphAF: a flow-based autoregressive model for molecular graph generation, 2020. URL https://arxiv.org/abs/2001.09382.

Martin Simonovsky and Nikos Komodakis. GraphVAE: Towards generation of small graphs using variational autoencoders, 2018. URL https://arxiv.org/abs/1802.03480.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

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

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3D molecule generation. *Advances in Neural Information Processing Systems*, 36, 2023.

Behrooz Tahmasebi and Stefanie Jegelka. Generalization bounds for canonicalization: A comparative study with group averaging. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=n0lXaskyk5.

Behrooz Tahmasebi and Stefanie Jegelka. Regularity in canonicalized models: A theoretical perspective. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025b.

Qingwen Tian, Yuxin Xu, Yixuan Yang, Zhen Wang, Ziqi Liu, Pengju Yan, and Xiaolin Li. EquiFlow: Equivariant conditional flow matching with optimal transport for 3D molecular conformation prediction, 2024. URL https://arxiv.org/abs/2412.11082.

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, 2024. URL https://arxiv.org/abs/2302.00482.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. DiGress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

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

Chenyu Wang, Cai Zhou, Sharut Gupta, Johnson Lin, Stefanie Jegelka, Stephen Bates, and Tommi Jaakkola. Learning diffusion models with flexible representation guidance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=cIGfKdfy3N.

Jike Wang, Hao Luo, Rui Qin, Mingyang Wang, Meijing Fang, Odin Zhang, Qiaolin Gou, Qun Su, Chao Shen, Ziyi You, Xiaozhe Wan, Liwei Liu, Chang-Yu Hsieh, Tingjun Hou, and Yu Kang. 3D molecular pocket-based generation with token-only large language model. *ChemRxiv*, 2024(0819), 2024. doi: 10.26434/chemrxiv-2024-0ckgt-v2. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2024-0ckgt-v2.

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

Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. Diffusion-based molecule generation with informative prior bridges. *Advances in Neural Information Processing Systems*, 35:36533–36545, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

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

Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3D molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

Yixian Xu, Yusong Wang, Shengjie Luo, Kaiyuan Gao, Tianyu He, Di He, and Chang Liu. Quotient-space diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3JPAkwSVc4.

Qi Yan, Zhengyang Liang, Yang Song, Renjie Liao, and Lele Wang. SwinGNN: Rethinking permutation invariance in diffusion models for graph generation. *arXiv preprint arXiv:2307.01646*, 2023.

Jiaxuan You, Jonathan M. Gomes-Selman, Rex Ying, and Jure Leskovec. Identity-aware graph neural networks. In *AAAI Conference on Artificial Intelligence*, 2021.

Yuning You, Ruida Zhou, Jiwoong Park, Haotian Xu, Chao Tian, Zhangyang Wang, and Yang Shen. Latent 3D graph diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

Leo Zhang, Kianoosh Ashouritaklimi, Yee Whye Teh, and Rob Cornish. SymDiff: Equivariant diffusion via stochastic symmetrisation. *arXiv preprint arXiv:2410.06262*, 2024.

Muhan Zhang, Pan Li, Yinglong Xia, Kai Wang, and Long Jin. Labeling trick: A theory of using graph neural networks for multi-node representation learning. *Advances in Neural Information Processing Systems*, 34:9061–9073, 2021.

Lingxiao Zhao, Xueying Ding, and Leman Akoglu. PARD: Permutation-invariant autoregressive diffusion for graph generation. *Advances in Neural Information Processing Systems*, 37:7156–7184, 2024.

Cai Zhou, Xiyuan Wang, and Muhan Zhang. From relational pooling to subgraph GNNs: A universal framework for more expressive graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42742–42768. PMLR, 2023a.

Cai Zhou, Xiyuan Wang, and Muhan Zhang. Facilitating graph neural networks with random walk on simplicial complexes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16172–16206, 2023b.

Cai Zhou, Xiyuan Wang, and Muhan Zhang. Unifying generation and prediction on graphs with latent graph diffusion. *Advances in Neural Information Processing Systems*, 37, 2024a.

Cai Zhou, Rose Yu, and Yusu Wang. On the theoretical expressive power and the design space of higher-order graph transformers. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2179–2187. PMLR, 02–04 May 2024b.

A Notation Guide

This section summarizes the notation used in the theoretical development. We use tildes to denote quantities after canonicalization, i.e., variables living on the canonical slice. We do not repeat standard diffusion-model notation unless it is needed to distinguish ambient-space and slice-space objects.

Table 3: Spaces, group actions, and canonicalization notation.

Notation	Meaning
$\mathcal{M}$	Measurable space of structured objects. In molecular generation, an element is $\mathbf{Z} = (\mathbf{X}, \mathbf{H}, \mathbf{A})$.
$\mathcal{G}$	Symmetry group acting on $\mathcal{M}$; after centering, typically $\mathcal{G} = S_N \times SO(3)$.
$g \cdot \mathbf{Z}$	Action of a group element $g \in \mathcal{G}$ on an object $\mathbf{Z}$.
$g\#\mu$	Pushforward of a measure μ by the map $\mathbf{Z} \mapsto g \cdot \mathbf{Z}$.
$\mathcal{O}(\mathbf{Z})$	Orbit of $\mathbf{Z}$: $\mathcal{O}(\mathbf{Z}) = \{g \cdot \mathbf{Z} : g \in \mathcal{G}\}$.
$\mathcal{M}/\mathcal{G}$	Quotient space whose elements are group orbits.
$\mathrm{Stab}(\mathbf{Z})$	Stabilizer of $\mathbf{Z}$: $\{g \in \mathcal{G} : g \cdot \mathbf{Z} = \mathbf{Z}\}$. A free action means the stabilizer is trivial.
λ	Haar probability measure on the compact group $\mathcal{G}$; uniform on S_N and uniform over rotations on $SO(3)$.
Ψ	Canonicalization map selecting an orbit representative, with $\Psi(\mathbf{Z}) \in \mathcal{O}(\mathbf{Z})$ and $\Psi(g \cdot \mathbf{Z}) = \Psi(\mathbf{Z})$.
$S = \Psi(\mathcal{M})$	Canonical slice, i.e., the image of the canonicalization map.
κ	Gauge map or canonical frame satisfying $x = \kappa(x) \cdot \Psi(x)$ and $\kappa(g \cdot x) = g\kappa(x)$ when defined.
$\delta_{\mathbf{Z}}$	Dirac measure at $\mathbf{Z}$, used in the factorization of invariant measures.

Table 4: Distributional and transport notation used in the theory.

Notation	Meaning
μ	Target distribution on the ambient object space $\mathcal{M}$, usually assumed $\mathcal{G}$-invariant.
$\nu = \Psi_{\#}\mu$	Slice distribution induced by pushing μ forward through the canonicalizer Ψ.
p_0, p_1, p_t	Ambient data, prior, and intermediate/noised distributions, respectively.
q_0, q_1	Canonical-slice data distribution and slice prior distribution, respectively.
$q_1^\star$	Aligned slice prior, e.g., the moment-matched Gaussian approximation to q_0 in the aligned-prior discussion of Appendix C.2.
$\mathbf{Z}, \tilde{\mathbf{Z}}$	Ambient variable and its canonicalized slice representative.
$\mathbf{Z}_t, \tilde{\mathbf{Z}}_t$	Intermediate ambient state and intermediate slice state along a diffusion or flow path.
$G \sim \lambda$	Latent random symmetry element used to lift slice variables back to the ambient space.
γ	Coupling between ambient endpoints $(\mathbf{Z}_0, \mathbf{Z}_1)$.
$\tilde{\gamma}$	Coupling between slice endpoints $(\tilde{\mathbf{Z}}_0, \tilde{\mathbf{Z}}_1)$.
γ^{lift}	Group-aligned lift of a slice coupling, defined by $(\mathbf{Z}_0, \mathbf{Z}_1) = (G \cdot \tilde{\mathbf{Z}}_0, G \cdot \tilde{\mathbf{Z}}_1)$ with $G \sim \lambda$.
γ_{mix}	Product ambient coupling, typically $p_0 \otimes p_1$, recovered by the lifted product slice coupling under an isotropic prior.
$\otimes$	Product coupling or product measure.
Φ_t	Conditional interpolation path mapping endpoints to an intermediate state.
$u_t(\cdot \mid \mathbf{Z}_0, \mathbf{Z}_1)$	Microscopic conditional vector field associated with a chosen endpoint pair.
v_t	Marginal vector field, given by the conditional expectation of the microscopic field.

B Related Work

Symmetries in generative models. Many datasets of interest in generative modeling are naturally defined only up to group actions—e.g., rotations and translations in 3D, or permutations in sets and

Table 5: Proof-level operators and auxiliary quantities.

Notation	Meaning
$U = \mathbf{Z}_1 - \mathbf{Z}_0$	Ambient displacement in the linear-path flow-matching analysis.
$\Delta = \tilde{\mathbf{Z}}_1 - \tilde{\mathbf{Z}}_0$	Slice displacement in the canonicalized linear-path analysis.
$\mathrm{Var}(\cdot \mid \cdot)$	Scalar conditional variance (the trace of the conditional covariance for vector-valued targets); the irreducible flow-matching regression error is an expected conditional variance.
$\mathrm{Cov}(\cdot \mid \cdot)$	Conditional covariance, used in the Gaussian supplement for within-slice variance.
$w_m(x)$	Responsibility weight of the m-th group copy in the symmetry-mixture score formula.
$\varrho(g \mid z)$	Posterior distribution over the latent group element given the intermediate ambient state.
$m_g(z)$, $m(z)$	Group-conditioned and marginal conditional mean drifts used to quantify symmetry ambiguity.
$\mathcal{L}_t(v)$	Population squared loss for a candidate vector field at time t.
$v^\star$	Bayes regressor, i.e., the conditional mean vector field minimizing $\mathcal{L}_t(v)$.
$(g, g)_{\#}\gamma = \gamma$	Diagonal invariance of a coupling under simultaneous group action on both endpoints.
$\mathcal{H}_{\mathrm{eq}}$	Class of equivariant functions or vector fields, used when comparing equivariant and canonicalized model classes.
$\mathcal{H}_{\mathrm{lift}}(\mathcal{A})$	Equivariant class induced by lifting functions from a slice function class $\mathcal{A}$.
$\mathrm{TV}(\cdot, \cdot)$	Total variation distance, used to compare induced model and target distributions.
$D_{\mathrm{KL}}(\cdot \Vert \cdot)$	Kullback-Leibler divergence, used to define the Gaussian projection for aligned priors.

graphs, so a principled generator should respect these symmetries to avoid spurious modes and over-counting equivalent configurations. A dominant approach is to build symmetry into the architecture by enforcing group equivariance (and thus invariance of the induced distribution), as in E(n)/SE(3)-equivariant message passing networks and their use in equivariant flows and diffusion/score-based models for geometric data (Satorras et al., 2021; Garcia Satorras et al., 2021; Hoogeboom et al., 2022). Systematic studies further clarify the design space and practical trade-offs of equivariant generative modeling under such geometric symmetries (Le et al., 2023; Lawrence et al., 2025b;a). An alternative architecture-agnostic perspective achieves symmetry by stochastic symmetrization (Zhang et al., 2024), which can enable the use of expressive and scalable non-equivariant backbones; learned canonicalization has been shown to provide a general mechanism for constructing equivariant functions from canonical representatives Kaba et al. (2023). However, canonicalization is not "free": for common groups there are fundamental continuity/stability obstructions, motivating softened or averaged constructions such as weighted frames to mitigate discontinuities near symmetric configurations (Dym et al., 2024; Lin et al., 2024; Ma et al., 2024). Related theory analyzes when canonicalization or group averaging offers better statistical behavior, highlighting distinct generalization regimes Tahmasebi & Jegelka (2025a); Puny et al. (2021); Lippmann et al. (2024); Tahmasebi & Jegelka (2025b); Duval et al. (2023). Complementing these perspectives, recent work in graph diffusion argues that enforcing strict invariance can make learning harder by inducing mixture-like objectives over symmetry transformations, and proposes post-hoc group randomization at sampling time to recover invariance without constraining the training model Yan et al. (2023). These works motivate the broader question of whether symmetry should be imposed architecturally or removed before learning.

Molecular generative models. Early work on discrete molecular structures includes latent-variable models (Jin et al., 2019; Simonovsky & Komodakis, 2018) and hierarchical graph generation (Jang et al., 2023). A major recent line uses diffusion processes on discrete graph attributes. Discrete denoising diffusion models operate directly on categorical attributes and have demonstrated strong performance at scale on molecular benchmarks (Vignac et al., 2022). In parallel, flow-based methods provide alternative likelihood-based or transport-inspired formulations for molecular graph generation (Luo et al., 2021; Shi et al., 2020), and discrete flow matching further improves sampling flexibility and efficiency while retaining strong generation quality (Qin et al., 2024; Hou

et al., 2024; Lee et al., 2025). Other diffusion approaches exploit degree information or spectral representations (Chen et al., 2023; Luo et al., 2023).

Generating molecules directly in 3D space has been explored with sequential and autoregressive schemes that place atoms step-by-step while maintaining geometric consistency (Gebauer et al., 2019; Luo & Ji, 2022; Feng et al., 2023; Wang et al., 2024). Diffusion models have become a dominant paradigm for 3D molecular generation by learning to denoise corrupted coordinates with architectures designed to respect Euclidean symmetries (Hoogeboom et al., 2022). Closely related conditional settings, such as conformer generation given a fixed molecular graph, have also benefited from geometric diffusion and flow matching formulations (Xu et al., 2022; Hassan et al., 2024; Hong et al., 2025; Wu et al., 2022).

A growing body of work aims to jointly generate discrete molecular graphs and continuous 3D geometries to avoid brittle post-hoc bond inference and to better couple chemical structure with spatial arrangement. MiDi (Vignac et al., 2023) proposes a mixed discrete–continuous diffusion approach that generates molecular graphs together with conformers in an end-to-end differentiable manner. Subsequent efforts further explore joint diffusion over comprehensive molecular representations (Huang et al., 2023; Hua et al., 2024; Irwin et al., 2024).

There are also latent generative models developed for 2D, 3D, or joint representations of molecules, including GeoLDM (Xu et al., 2023), LGD (Zhou et al., 2024a), and LDM-3DG (You et al., 2024). More recently, GeoRCG (Li et al., 2025) further utilizes a two-stage generation: first generating molecule representations, then using geometric representations to guide the molecule generation. By contrast, REED (Wang et al., 2025) leverages the pretrained molecular representation as the guidance, aligning internal diffusion model features with these high-quality representations to accelerate training and improve generation performance. We emphasize that our analysis and proposed methods are also applicable and effective for these latent/representation space generative models under mild conditions.

Non-equivariant, alignment-based, and quotient-space biochemical generators. A related line of work in molecular and biomolecular generation relaxes fully equivariant architectures and instead handles rigid-motion symmetries through alignment, augmentation, symmetrization, or quotient-space modeling. In 3D molecule and conformer generation, geometric diffusion models often align coordinates or targets when defining losses or sampling procedures to reduce redundant rotational degrees of freedom (Xu et al., 2022), and recent work explicitly studies scalable non-equivariant 3D molecule generation via rotational alignment (Ding & Hofmann, 2025). Similar ideas also appear in biomolecular structure modeling and design, where protein and complex generation systems rely on geometric alignment, frame choices, or symmetry-aware losses to stabilize learning over equivalent structures (Watson et al., 2023; Corso et al., 2023; Abramson et al., 2024; Jing et al., 2024). Architecture-agnostic stochastic symmetrization SymDiff further shows that equivariance can be recovered without hard-coding it into the backbone (Zhang et al., 2024). However, it still requires an S_N-equivariant backbone, only relaxing $O(3)$ – our approach relaxes both S_N and $SE(3)$, enabling fully non-equivariant architectures with positional encodings. Most closely related to our motivation, Quotient-Space Diffusion Models formulate diffusion directly on the quotient space to avoid learning redundant symmetry directions and to provide a principled sampler, in contrast to heuristic alignment strategies (Xu et al., 2026). Our work shares this quotient-space view but takes a different route: rather than defining the diffusion intrinsically on the quotient, we choose a measurable canonical slice, train a flexible non-equivariant model on the slice, and recover the invariant distribution by Haar randomization. This canonicalization perspective gives an explicit slice distribution, a conditional-variance explanation for training acceleration, and a practical mechanism for combining non-equivariant backbones with aligned priors and optimal transport.

Optimal transport. In continuous-time generative modeling, optimal transport (OT) is often used as a principle for selecting cost-effective and more structured probability paths between a simple base distribution and the data distribution. A widely used choice in flow matching is the OT displacement interpolation, which yields trajectories closer to minimal-cost transport and empirically leads to faster convergence and fewer sampling steps (Lipman et al., 2022). Related "trajectory straightening" viewpoints, such as Rectified Flow, can be interpreted as progressively transforming the learned dynamics toward straighter transport-like paths, further improving numerical stability and reducing the number of integration steps required at inference (Liu et al., 2022; Tong et al., 2024; Kornilov et al., 2024).

In molecular generation, OT-inspired objectives have been adopted more explicitly to enable high-quality few-step generation under geometric symmetries. For 2D molecular graph generation, MolGAN (Cao & Kipf, 2022) was among the first to successfully use the Wasserstein-1 distance to stabilize the training of molecular graph generators. Subsequent work such as BWFlow (Jiang et al., 2025) also demonstrates the effectiveness of OT in 2D molecule design. For 3D molecule generation, various methods have demonstrated that OT can serve not only as a theoretical lens but also as a practical design tool for fast, high-fidelity molecule synthesis (Song et al., 2023; Tian et al., 2024; Hong et al., 2025).

C Omitted Proofs

This section provides all omitted proofs and more detailed discussions on our theoretical results, presented mainly in Section 3.

C.1 Proof for Section 3.1

This subsection proves that canonicalized generative models can induce universal invariant distributions and outperform those equivariant baselines trained directly on invariant data.

C.1.1 Proof for Universality of Canonical Parameterizations

We first show the correctness and sufficiency of canonical slice modeling.

Lemma C.1 (Factorization of invariant measures; Lemma 3.1 in main text)**.** *Suppose Assumptions 2.5 and 2.6 hold. Let μ be any $\mathcal{G}$-invariant probability measure on $\mathcal{M}$. Let Ψ be an orbit representative map defined μ-a.s., and let $\nu = \Psi\#\mu$ be the slice distribution on $S = \Psi(\mathcal{M})$. Then*

$$\mu \;=\; \int_S \left(\int_{\mathcal{G}} \delta_{g\cdot \mathbf{Z}} \, \mathrm{d}\lambda(g) \right) \mathrm{d}\nu(\mathbf{Z}). \tag{16}$$

Equivalently, if $\tilde{\mathbf{Z}} \sim \nu$, $g \sim \lambda$ independent, then $g \cdot \tilde{\mathbf{Z}} \sim \mu$.

Proof. This is a known result, included for completeness. Let $\mathbf{Z} \sim \mu$ and define $\tilde{\mathbf{Z}} = \Psi(\mathbf{Z}) \in S$. By definition of orbit representative, there exists $g(\mathbf{Z}) \in \mathcal{G}$ with $\mathbf{Z} = g(\mathbf{Z}) \cdot \tilde{\mathbf{Z}}$. Under 2.5, $g(\mathbf{Z})$ is unique a.s. Fix any measurable $A \subseteq \mathcal{G}$. For any $h \in \mathcal{G}$, invariance of μ implies

$$\mathbb{P}(g(\mathbf{Z}) \in A \mid \tilde{\mathbf{Z}}) = \mathbb{P}(h\, g(\mathbf{Z}) \in A \mid \tilde{\mathbf{Z}}),$$

so the conditional law of $g(\mathbf{Z})$ given $\tilde{\mathbf{Z}}$ is left-invariant. By uniqueness of Haar probability on compact $\mathcal{G}$, this conditional law is λ. Therefore,

$$\begin{aligned} \mu(B) = \mathbb{E}[\mathbf{1}\{g(\mathbf{Z}) \cdot \tilde{\mathbf{Z}} \in B\}] &= \mathbb{E}\Big[\int_{\mathcal{G}} \mathbf{1}\{g \cdot \tilde{\mathbf{Z}} \in B\} \, \mathrm{d}\lambda(g) \Big] \\ &= \int_S \Big(\int_{\mathcal{G}} \delta_{g\cdot \mathbf{Z}}(B) \, \mathrm{d}\lambda(g) \Big) \mathrm{d}\nu(\mathbf{Z}), \end{aligned} \tag{17}$$

which is equation 4. □

Corollary C.2 (Sufficiency of slice modeling; Corollary 3.2 in main text)**.** *To model any invariant target μ, it suffices to model the slice distribution ν; invariance is recovered by Haar randomization.*

In the next part, we show the universality of canonicalized parameterizations over invariant and equivariant functions.

Proposition C.3 (Universality over canonicalized parameterizations; Proposition 3.3 in the main text)**.** *Let $\mathcal{G}$ act continuously and orthogonally on a compact set $K \subset \mathbb{R}^d$. Suppose we have a (measurable) canonicalization map $\Psi : K \to K$ as Definition 2.3, and a (measurable) gauge map $\kappa : K \to \mathcal{G}$ s.t.*

$$\Psi(g \cdot x) = \Psi(x), \; \kappa(g \cdot x) = g\kappa(x), \; x = \kappa(x) \cdot \Psi(x) \tag{18}$$

(defined on a full-measure set; under free actions, κ is unique a.s.). Consider the parametrization

$$\phi(x) = \kappa(x) \cdot f(\Psi(x)), \tag{19}$$

where f is a universal approximator on $\Psi(K)$. Then ϕ is a universal approximator of continuous $\mathcal{G}$-equivariant functions on K, and $f \circ \Psi$ is universal for continuous $\mathcal{G}$-invariant functions on K.

Proof. Let $F : K \to \mathbb{R}^d$ be continuous and $\mathcal{G}$-equivariant: $F(g \cdot x) = g \cdot F(x)$. Define its restriction to the slice $S := \Psi(K)$ by $F_S(z) := F(z)$ for $z \in S$. For any $x \in K$, write $x = \kappa(x) \cdot \Psi(x)$. By equivariance,

$$F(x) = F(\kappa(x) \cdot \Psi(x)) = \kappa(x) \cdot F(\Psi(x)) = \kappa(x) \cdot F_S(\Psi(x)). \tag{20}$$

Since f is universal on S, for any $\varepsilon > 0$ there exists f such that $\sup_{z \in S} \|f(z) - F_S(z)\| < \varepsilon$. Then

$$\sup_{x \in K} \|\phi(x) - F(x)\| \leq \sup_{z \in S} \|f(z) - F_S(z)\| < \varepsilon, \tag{21}$$

using that the group action is norm-preserving. □

C.1.2 Proof for Stronger Expressivity via Symmetry Breaking

We now show the superiority of canonicalized generative models with expressive (possibly) non-equivariant networks over standard generative models with equivariant backbones.

Expressivity gain in S_N from canonical ordering. It is known that the expressivity of message-passing based GNNs is bounded by the first-order Weisfeiler–Leman (1-WL) algorithm on abstract graphs in terms of distinguishing non-isomorphic graphs (Xu et al., 2018). This immediately implies their limitations in counting substructures such as cycles (Huang et al., 2022) and even link prediction (Zhang et al., 2021). Analogous issues exist for transformers due to their equivalence with MPNNs (Cai et al., 2023), and equivariant GNNs including EGNNs (Satorras et al., 2021) when processing 3D geometric graphs (Li et al., 2023; 2024). There are some mainstream methods to overcome the permutation-related expressivity pitfalls:

1. *High-order architectures*: high-order GNNs (Morris et al., 2019; 2020; Maron et al., 2019) and high-order transformers (Kim et al., 2022; Zhou et al., 2024b) have provably more powerful expressivity within and beyond the k-WL framework, at the cost of exponentially higher computation; for example, SwinGNN (Yan et al., 2023) adopts 3-WL equivalent PPGN (Maron et al., 2019) as the denoising backbone. We argue, however, stronger expressivity does not need to emerge from high-order networks with exponential complexity: symmetry breaking also improves expressivity as detailed as follows.
2. *Symmetry breaking*: by introducing asymmetry intentionally, the expressivity of underlying network can be provably enhanced. A widely adopted technique is "labeling" the nodes with extra identity features (You et al., 2021; Zhang et al., 2021; Huang et al., 2022; Zhou et al., 2023a), and the expressivity monotonically increases with the number of IDs. Unfortunately, these methods need the relational pooling or averaging to produce invariant outputs, which again introduces complexity at the scale of order of the group - canonicalization offers a way to produce appropriate and stable outputs (Dong et al., 2024) without the costly enumeration and averaging procedure.
3. *Positional encodings and structural encodings*: graph structure-based PEs significantly improve the expressivity and practical performance of neural backbones (Rampasek et al., 2022). Popular PEs include spatial structure-based (e.g. RWSE (Dwivedi et al., 2021) and RRWP (Ma et al., 2023b)) and spectra-based (using eigenvalues and eigenvectors of graph Laplacians (Ma et al., 2023a; Lim et al., 2022; Bo et al., 2023) or Hodge Laplacians (Zhou et al., 2023b)). However, different from the fixed positional encodings in images or texts based on absolute indices of tokens, the calculation of the above PEs on graphs rely on ground-truth structures, which are not available in diffusion generation. Some generative models like DeFoG (Qin et al., 2024) utilize an estimated PE from the current noisy graph, which we argue might not be the optimal choice. In comparison, our canonicalization rank can also be viewed as spatial and spectral PEs, yet yields almost no computation overhead and no train-test gap.

The three observations above are specific to the S_N side of the symmetry: 1-WL limitations, node IDs, and graph positional encodings all target the loss of information caused by quotienting over vertex permutations. In this sense, canonical ordering provides a stable symmetry-breaking mechanism that strengthens the effective hypothesis class on the permutation quotient.

Expressivity gain in $SO(3)$ from measurable gauge fixing. The rotational side gives a different, complementary separation. Practical equivariant architectures built from continuous layers realize continuous equivariant maps, whereas canonicalization may use a measurable gauge that is discontinuous at unavoidable frame-switching sets. This allows canonicalized non-equivariant models to represent equivariant targets outside the continuous equivariant class.

Proposition C.4 (Gauge lifting and strict separation for $SO(3)$)**.** *Let $\mathcal{X} \subset \mathbb{R}^{N\times 3}$ be a space of centered point clouds with the $SO(3)$ action $R \cdot X := XR^\top$. Let $\mathcal{X}_0 \subseteq \mathcal{X}$ be a full-measure subset on which the action is free. Suppose there exist measurable maps $\Psi : \mathcal{X}_0 \to S$ and $\kappa : \mathcal{X}_0 \to SO(3)$ such that, for all $R \in SO(3)$ and $X \in \mathcal{X}_0$,*

$$\Psi(R \cdot X) = \Psi(X), \qquad \kappa(R \cdot X) = R\kappa(X), \qquad X = \kappa(X) \cdot \Psi(X). \tag{22}$$

For any measurable $a : S \to SO(3)$, define

$$F_a(X) := \kappa(X)a(\Psi(X)). \tag{23}$$

Then F_a is $SO(3)$-equivariant on $\mathcal{X}_0$. Moreover, every continuous $SO(3)$-equivariant map $f : \mathcal{X} \to SO(3)$ is represented by this form with $a = f|_S$. If κ is not equal on $\mathcal{X}_0$ to the restriction of any continuous $SO(3)$-equivariant map on $\mathcal{X}$, then the canonicalized class strictly contains the continuous equivariant class.

Proof. Using equation 22, for any $R \in SO(3)$,

$$F_a(R \cdot X) = \kappa(R \cdot X)a(\Psi(R \cdot X)) = R\kappa(X)a(\Psi(X)) = RF_a(X), \tag{24}$$

so F_a is equivariant. For any continuous equivariant f, choose $a = f|_S$. Since $X = \kappa(X) \cdot \Psi(X)$,

$$f(X) = f(\kappa(X) \cdot \Psi(X)) = \kappa(X)f(\Psi(X)) = F_a(X). \tag{25}$$

For strictness, choose $a \equiv I$, so $F_a = \kappa$. If κ is not the restriction of any continuous equivariant map, then F_a lies in the canonicalized class but not in the continuous equivariant class. This obstruction is consistent with the impossibility of globally continuous canonical frames for groups such as $SO(d)$ (Dym et al., 2024). □

Remark C.5. Proposition C.4 is a function-class statement, not a continuity claim about the learned neural backbone. The non-equivariant backbone only needs to model functions on the canonical slice, while the possibly discontinuous gauge is supplied by the canonicalizer. Stabilizers or degenerate frames can be handled by restricting to the full-measure free-action set, as in Assumption 2.5.

TV guarantee for canonicalized non-equivariant backbones. The preceding S_N and $SO(3)$ discussions justify why the slice model class induced by canonicalization can be at least as expressive as an equivariant ambient baseline. We next show that any such slice-level advantage transfers to the invariant target distribution after Haar randomization.

Theorem C.6 (Canonicalized models match (or improve) compact-group equivariant baselines in TV)**.** *Under Assumptions 2.5 and 2.6, let $\mathcal{G}$ be the compact symmetry group after centering, e.g., S_N, $SO(3)$, or $S_N \times SO(3)$. Let μ be any $\mathcal{G}$-invariant target distribution on $\mathcal{M}$ and let $\nu := \Psi_{\#}\mu$ be its canonical-slice distribution on $S = \Psi(\mathcal{M})$. Assume there exists an* equivariant baseline *(defined in ambient space) producing an invariant model distribution $\hat{\mu}_{\theta(\mathrm{eq})}$ such that*

$$\mathrm{TV}\big(\hat{\mu}_{\theta(\mathrm{eq})}, \mu\big) < \varepsilon. \tag{26}$$

Let $\hat{\nu}_{\theta(\mathrm{eq})} := \Psi_{\#}\hat{\mu}_{\theta(\mathrm{eq})}$ be the induced slice distribution. If the slice model class $\{\hat{\nu}_{\theta(\mathrm{free})}\}$ is at least as expressive on S in the sense that

$$\inf_{\theta(\mathrm{free})} \mathrm{TV}\big(\hat{\nu}_{\theta(\mathrm{free})}, \nu\big) \;\le\; \mathrm{TV}\big(\hat{\nu}_{\theta(\mathrm{eq})}, \nu\big), \tag{27}$$

then there exists $\theta(\mathrm{free})$ such that the Haar-randomized model

$$\hat{\mu}_{\theta(\mathrm{free})} := \int_S \left(\int_{\mathcal{G}} \delta_{g\cdot\mathbf{Z}} \, \mathrm{d}\lambda(g) \right) \mathrm{d}\hat{\nu}_{\theta(\mathrm{free})}(\mathbf{Z}) \tag{28}$$

satisfies $\mathrm{TV}\big(\hat{\mu}_{\theta(\mathrm{free})}, \mu\big) < \varepsilon$.

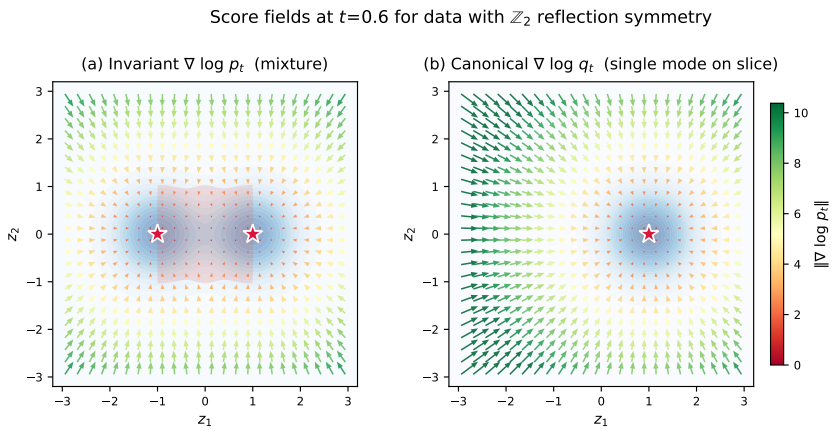


Figure 4: Toy illustration of score fields at intermediate diffusion time $t = 0.6$ for a $\mathbb{Z}_2$-symmetric Gaussian mixture in 2D. (a) The invariant score $\nabla \log p_t$ is a responsibility-weighted mixture of per-mode scores (Eq. (7)). Near the symmetry axis, the two modes' contributions cancel, creating a low-magnitude "dead zone" (red shading) where the score provides little learning signal and conflicting gradient directions. (b) On the canonical slice, only a single mode remains. The score $\nabla \log q_t$ is smooth, unambiguous, and consistently points toward the mode center, eliminating the mixture complexity entirely. Arrow length is proportional to score magnitude; color encodes $|\nabla \log p_t|$ (green = high, red = low).

Proof. First, pushforward cannot increase TV (data processing inequality). Hence

$$\mathrm{TV}(\hat{\nu}_{\theta(\mathrm{eq})}, \nu) = \mathrm{TV}\big(\Psi_{\#}\hat{\mu}_{\theta(\mathrm{eq})}, \Psi_{\#}\mu\big) \le \mathrm{TV}\big(\hat{\mu}_{\theta(\mathrm{eq})}, \mu\big) < \varepsilon. \tag{29}$$

By equation 27, the infimum of the slice-model error is strictly below ε. Hence, there exists $\theta(\mathrm{free})$ such that $\mathrm{TV}(\hat{\nu}_{\theta(\mathrm{free})}, \nu) < \varepsilon$. Finally, Haar randomization is a Markov kernel T (cf. Lemma 3.1), and thus is TV-nonexpansive:

$$\mathrm{TV}\big(T(\hat{\nu}_{\theta(\mathrm{free})}), T(\nu)\big) \le \mathrm{TV}\big(\hat{\nu}_{\theta(\mathrm{free})}, \nu\big). \tag{30}$$

Since $T(\nu) = \mu$ (Lemma 3.1), we conclude $\mathrm{TV}(\hat{\mu}_{\theta(\mathrm{free})}, \mu) < \varepsilon$. □

C.2 Proof for Section 3.2

In this subsection, we provide more detailed analysis and complete proofs for our results regarding accelerating diffusion and flow model training through canonicalization.

C.2.1 Proof for Mixture Structure in Diffusion Score

This part mainly considers score-based diffusion modeling, showing that the existence of group symmetry in invariant data induces the mixture structure of the score function. A toy visualization of the corresponding intermediate-time score field is provided in Figure 4, which illustrates the mixture-induced ambiguity near the symmetry axis and the simplification brought by canonicalization.

Proposition C.7 (Score of a symmetry mixture)**.** *Assume q is differentiable and positive where needed. Then*

$$\nabla \log p(x) = \sum_{m=1}^{M} w_m(x)\, g_m \cdot \nabla \log q(g_m^{-1} \cdot x), \quad w_m(x) := \frac{q(g_m^{-1} \cdot x)}{\sum_{j=1}^{M} q(g_j^{-1} \cdot x)}. \tag{31}$$

Proof. Differentiate $\log p(x) = \log\big(\frac{1}{M}\sum_m q(g_m^{-1} \cdot x)\big)$ and apply the chain rule. Orthogonality implies $\nabla_x q(g_m^{-1} \cdot x) = g_m \cdot \nabla q(g_m^{-1} \cdot x)$. □

C.2.2 Proof for Conditional Flow Variance Reduction

This part highlights our central result: canonical training reduces conditional variance in flow-matching.

Irreducible error in flow matching. First we prove a standard lemma, stating that the irreducible flow-matching error is the conditional variance.

Lemma C.8 (Irreducible flow-matching error = conditional variance)**.** *Let $\hat{v}$ be any measurable predictor of $v_t(\mathbf{Z}_t)$ from $(t, \mathbf{Z}_t)$. Then the minimum achievable MSE satisfies*

$$\inf_{\hat{v}} \mathbb{E}\big[\, \|\hat{v}(t, \mathbf{Z}_t) - (\mathbf{Z}_1 - \mathbf{Z}_0)\|^2 \,\big] = \mathbb{E}\big[\mathrm{Var}(\mathbf{Z}_1 - \mathbf{Z}_0 \mid t, \mathbf{Z}_t)\big]. \tag{32}$$

Proof. This is the standard L^2 regression identity: the conditional expectation $\mathbb{E}[\mathbf{Z}_1 - \mathbf{Z}_0 \mid t, \mathbf{Z}_t]$ is the unique minimizer, and the minimum risk equals the expected conditional variance (law of total variance). □

Corollary C.9 (Deterministic couplings yield zero irreducible error)**.** *If $\mathbf{Z}_1 - \mathbf{Z}_0$ is (a.s.) a deterministic function of $(t, \mathbf{Z}_t)$, then the right-hand side is 0, and in principle the flow can be learned with arbitrarily small error (up to approximation and optimization).*

Below we give a concrete example of this irreducible error, showing that the minimum achievable MSE loss is strictly positive in the presence of multi-group ambiguity.

Quantifying symmetry ambiguity through posterior collision bounds We now strengthen the variance-decomposition viewpoint by (i) giving an exact identity for the *symmetry-ambiguity* term in Theorem 3.5, (ii) deriving lower bounds purely in terms of the posterior $\varrho(g \mid Z_t)$ over the latent symmetry.

Under the group-aligned lifted coupling, write the posterior over the latent symmetry as $\varrho(g \mid z) := \mathbb{P}(G = g \mid Z_t = z)$ (finite group case; for compact groups interpret $\varrho(\cdot \mid z)$ as a posterior density w.r.t. Haar).

We first show that the ambiguity term can be written as a pairwise mean-separation identity. The following identity is standard (*variance via i.i.d. copies*), but its specialization here yields an explicit geometric handle on the symmetry-ambiguity term.

Lemma C.10 (Symmetry ambiguity as pairwise conditional-mean separation)**.** *Define the group-conditioned conditional mean drift*

$$m_g(z) := \mathbb{E}[U \mid Z_t = z,\ G = g], \qquad m(z) := \mathbb{E}[U \mid Z_t = z]. \tag{33}$$

Then the ambiguity term in Theorem 3.5 satisfies, for each z,

$$\mathrm{Var}(\mathbb{E}[U \mid Z_t, G] \mid Z_t = z) = \mathbb{E}_{g \sim \varrho(\cdot|z)}\big[\|m_g(z) - m(z)\|^2\big] \tag{34}$$

$$= \frac{1}{2}\, \mathbb{E}_{g, g' \overset{i.i.d.}{\sim} \varrho(\cdot|z)}\big[\|m_g(z) - m_{g'}(z)\|^2\big]. \tag{35}$$

Proof. Eq. (34) is the definition of conditional variance of a discrete random variable taking values $m_G(z)$. Eq. (35) follows from the standard identity $\mathrm{Var}(X) = \frac{1}{2}\mathbb{E}\|X - X'\|^2$ for i.i.d. copies (X, X') (applied conditionally on $Z_t = z$). □

We hereby provide the lower bounds from posterior uncertainty (collision probability). Lemma C.10 implies that ambiguity is large whenever the posterior $\varrho(g \mid Z_t)$ spreads mass over group elements with meaningfully different conditional mean drifts.

Proposition C.11 (Posterior-collision lower bound)**.** *Fix z and let $\mathcal{G}_\Delta(z) \subseteq \mathcal{G}$ be any subset such that for all $g \in \mathcal{G}_\Delta(z)$,*

$$\|m_g(z) - m(z)\| \geq \Delta(z). \tag{36}$$

Then

$$\mathrm{Var}(\mathbb{E}[U \mid Z_t, G] \mid Z_t = z) \ \geq\ \Delta(z)^2\, \mathbb{P}\big(G \in \mathcal{G}_\Delta(z) \mid Z_t = z\big). \tag{37}$$

In particular, for a finite group, if $\|m_g(z) - m_{g'}(z)\| \geq \Delta(z)$ for every pair of distinct elements g, g' in the posterior support, then

$$\mathrm{Var}(\mathbb{E}[U \mid Z_t, G] \mid Z_t = z) \ \geq\ \frac{\Delta(z)^2}{2}\, \Big(1 - \sum_g \varrho(g \mid z)^2\Big), \tag{38}$$

where $1 - \sum_g \varrho(g \mid z)^2$ is the complement of the posterior collision probability.

Proof. The first bound is immediate from equation 34 by keeping only $g \in \mathcal{G}_\Delta(z)$ and using $\|m_g(z) - m(z)\|^2 \geq \Delta(z)^2$ there. For the second bound, use equation 35 and lower-bound the pairwise distance by $\Delta(z)$ whenever $g \neq g'$ in the posterior support. Then

$$\mathbb{P}(g \neq g' \mid z) = 1 - \sum_g \varrho(g \mid z)^2, \tag{39}$$

and the stated inequality follows, including the factor $1/2$ from equation 35. □

The collision term $1 - \sum_g \varrho(g \mid Z_t)^2$ is 0 iff G is a.s. determined by Z_t. Thus, at a state where the posterior assigns positive mass to distinct conditional mean drifts, symmetry ambiguity creates a strictly positive variance component.

Discussions on lifted coupling. A subtle but important point is that the group-aligned lift in equation 10 shares the same latent G across (Z_0, Z_1), which *a priori* could introduce dependence. However, for *standard isotropic Gaussian* priors and orthogonal group actions (including S_N permutations and $SO(3)$ rotations), this dependence disappears.

Proposition C.12 (Lift of a product slice coupling equals the ambient product coupling)**.** *Assume $\mathcal{G}$ acts orthogonally on $\mathbb{R}^d$. Let $G \sim \lambda$ and $(\widetilde{Z}_0, \widetilde{Z}_1) \sim q_0 \otimes \mathcal{N}(0, I)$ be independent, and define $(Z_0, Z_1) = (G \cdot \widetilde{Z}_0,\ G \cdot \widetilde{Z}_1)$. Then (i) $Z_1 \sim \mathcal{N}(0, I)$ and $Z_1 \perp G$, and (ii) $Z_1 \perp Z_0$. Moreover, if the invariant data distribution satisfies the disintegration $p_0 = \int (g\cdot)_{\#} q_0 \, d\lambda(g)$, then the lifted ambient coupling satisfies*

$$(Z_0, Z_1) \sim p_0 \otimes \mathcal{N}(0, I). \tag{40}$$

Proof. For any measurable B and any g, orthogonal invariance of $\mathcal{N}(0, I)$ gives $\mathbb{P}(Z_1 \in B \mid G = g) = \mathbb{P}(g \cdot \widetilde{Z}_1 \in B) = \mathbb{P}(\widetilde{Z}_1 \in B)$, which does not depend on g, hence $Z_1 \perp G$ and $Z_1 \sim \mathcal{N}(0, I)$. For any measurable A, B,

$$\mathbb{P}(Z_0 \in A, Z_1 \in B) = \mathbb{E}\Big[\mathbf{1}_{\{Z_0 \in A\}} \mathbb{P}(Z_1 \in B \mid G, \widetilde{Z}_0)\Big] = \mathbb{P}(\widetilde{Z}_1 \in B)\, \mathbb{P}(Z_0 \in A), \tag{41}$$

since $\widetilde{Z}_1$ is independent of $(G, \widetilde{Z}_0)$ and $\mathbb{P}(g \cdot \widetilde{Z}_1 \in B)$ is constant in g. Thus $Z_1 \perp Z_0$. If additionally $Z_0 \sim p_0$ via the stated disintegration, the joint law is $p_0 \otimes \mathcal{N}(0, I)$. □

Therefore, the commonly used ambient product coupling $\gamma_{\text{mix}} = p_0 \otimes \mathcal{N}(0, I)$ can be viewed as a special case of slice training with product coupling and isotropic Gaussian slice prior, thus *conditional variances of two paradigms become directly comparable*. Canonicalization does *not* change the coupling in this case; it changes the *representation* so that the group ambiguity term in Theorem 3.5 becomes explicit and avoidable on the slice. Hence our theory covers most practical training regimes without optimal transport.

However, this generally does not hold if $\tilde{Z}_1$ is group-aware (e.g., learned or group-statistics related prior in our implementation), or $\tilde{Z}_1$ and $\tilde{Z}_0$ are not independent (e.g., optimal transport on the canonicalized slice). In particular, if the slice prior is *not* isotropic (e.g., an aligned Gaussian $\mathcal{N}(\mu, \Sigma)$ with $g\Sigma g^\top \neq \Sigma$ for some g), then $g \cdot \widetilde{Z}_1$ depends on g in distribution, so Z_1 is no longer independent of G, and the lifted coupling generally is *not* a product. Likewise, if the slice coupling $\widetilde{\gamma}$ is non-product (e.g., OT/Monge), the lifted coupling remains non-product.

Canonicalization reduces conditional variance via symmetry ambiguity elimination. We now prove our central result on the advantages of canonicalized flow matching.

Theorem C.13 (Variance decomposition under group-aligned lift; Theorem 3.5 in main text)**.** *Assume $\mathcal{G}$ acts orthogonally. Under the group-aligned lifted coupling $G \sim \lambda$ independent of $(\widetilde{Z}_0, \widetilde{Z}_1)$ and $(Z_0, Z_1) = (G \cdot \widetilde{Z}_0, G \cdot \widetilde{Z}_1)$, let $Z_t = (1-t)Z_0 + tZ_1$ and $\widetilde{Z}_t = (1-t)\widetilde{Z}_0 + t\widetilde{Z}_1$. With $U := Z_1 - Z_0 = G \cdot \Delta$ and $\Delta := \widetilde{Z}_1 - \widetilde{Z}_0$, we have*

$$\mathrm{Var}(U \mid Z_t) = \underbrace{\mathbb{E}\Big[\mathrm{Var}(\Delta \mid \widetilde{Z}_t) \mid Z_t\Big]}_{\textit{within-slice difficulty}} + \underbrace{\mathrm{Var}(\mathbb{E}[U \mid Z_t, G] \mid Z_t)}_{\textit{symmetry ambiguity} \geq 0}. \tag{42}$$

Consequently,

$$\mathbb{E}\big[\mathrm{Var}(U \mid Z_t)\big] \;\geq\; \mathbb{E}\big[\mathrm{Var}(\Delta \mid \widetilde{Z}_t)\big]. \tag{43}$$

Proof. Apply the standard law of total variance with respect to the latent symmetry variable G:

$$\mathrm{Var}(U \mid Z_t) = \mathbb{E}[\mathrm{Var}(U \mid Z_t, G) \mid Z_t] + \mathrm{Var}(\mathbb{E}[U \mid Z_t, G] \mid Z_t)\,. \tag{44}$$

The second term is nonnegative. For the first term, conditioning on (Z_t, G) determines $\widetilde{Z}_t = G^{-1} \cdot Z_t$. Under the lifted coupling, $(\Delta, \widetilde{Z}_t)$ is independent of G, and orthogonality implies $\mathrm{Var}(G \cdot X \mid \cdot) = \mathrm{Var}(X \mid \cdot)$. Therefore,

$$\mathrm{Var}(U \mid Z_t, G) = \mathrm{Var}(G \cdot \Delta \mid \widetilde{Z}_t, G) = \mathrm{Var}(\Delta \mid \widetilde{Z}_t). \tag{45}$$

Substituting into equation 44 yields equation 11. Taking expectations over Z_t and using the tower property gives equation 12. □

The lower bound in Theorem 3.5 is the "within-slice" conditional variance; the gap is exactly the additional uncertainty induced by not knowing which symmetry element generated the observation. This formalizes the intuition that symmetry induces posterior multi-modality over group elements, which inflates the conditional variance and creates an irreducible error floor for coarse solvers.

Remark C.14 (What this does *and does not* compare). The theorem above compares "with vs. without marginalizing G" *under the same lifted coupling*. It does *not* imply that a canonicalized training paradigm universally dominates any other paradigm under arbitrary choices of couplings. Its value is to isolate a nonnegative variance component that arises purely from symmetry ambiguity and is absent on the slice. However, as discussed in Proposition C.12, the lifted coupling can easily recover the standard product with Gaussian noises used in most practical flow matching models (w/o OT).

In the next part we will further explain Remark 3.6. Fix a coupling γ on (Z_0, Z_1) and define the linear interpolation

$$Z_t = (1-t)Z_0 + tZ_1, \qquad U := Z_1 - Z_0. \tag{46}$$

For any predictor $v : \mathbb{R}^d \to \mathbb{R}^d$, the population squared loss at time t is

$$\mathcal{L}_t(v) := \mathbb{E}\big[\|v(Z_t) - U\|^2\big]. \tag{47}$$

It is standard that the Bayes regressor is $v^\star(z) = \mathbb{E}[U \mid Z_t = z]$ and the Bayes risk equals the conditional variance:

$$\inf_v \mathcal{L}_t(v) = \mathbb{E}\big[\mathrm{Var}(U \mid Z_t)\big]. \tag{48}$$

Canonicalization does not improve the ambient Bayes risk of equivariant models. Let a compact group $\mathcal{G}$ act orthogonally on $\mathbb{R}^d$ (e.g. permutations and rotations). Let γ be *diagonal-invariant*:

$$(g, g)_\# \gamma = \gamma \quad \forall g \in \mathcal{G}, \tag{49}$$

which holds in particular for the ambient product coupling $p_0 \otimes \mathcal{N}(0, I)$ when p_0 is $\mathcal{G}$-invariant, and also for symmetrized OT couplings.

Proposition C.15 (Equivariance of the Bayes regressor under diagonal invariance)**.** *Assume $\mathcal{G}$ acts orthogonally and γ is diagonal-invariant. Then the Bayes regressor $v^\star(z) = \mathbb{E}[U \mid Z_t = z]$ is $\mathcal{G}$-equivariant:*

$$v^\star(g \cdot z) = g \cdot v^\star(z) \qquad \forall g \in \mathcal{G}. \tag{50}$$

Consequently, restricting to the equivariant function class $\mathcal{H}_{\mathrm{eq}} := \{v :\ v(g \cdot z) = g \cdot v(z)\}$ does not *increase the minimal population loss:*

$$\inf_{v \in \mathcal{H}_{\mathrm{eq}}} \mathcal{L}_t(v) = \inf_v \mathcal{L}_t(v) = \mathbb{E}[\mathrm{Var}(U \mid Z_t)]. \tag{51}$$

Proof. Diagonal invariance implies $(Z_t, U) \overset{d}{=} (g \cdot Z_t, g \cdot U)$. Hence for any measurable set A,

$$\mathbb{E}[U \mathbf{1}_{\{Z_t \in A\}}] = \mathbb{E}[g \cdot U\ \mathbf{1}_{\{g \cdot Z_t \in A\}}] = g \cdot \mathbb{E}[U \mathbf{1}_{\{Z_t \in g^{-1} A\}}], \tag{52}$$

which yields $v^\star(g \cdot z) = g \cdot v^\star(z)$ by Radon–Nikodym characterization of conditional expectation. The equality of infima follows because $v^\star \in \mathcal{H}_{\mathrm{eq}}$. □

Remark C.16. If one keeps the *same ambient regression problem* equation 48, then canonicalizing the inputs does not improve the population optimum for an equivariant model class, because the Bayes regressor is already equivariant (Proposition C.15).

Canonical conditions expand the effective hypothesis class of non-equivariant models. Canonicalization is most useful when the model is *not* architecturally equivariant. Intuitively, a non-equivariant model in ambient space must learn symmetry averaging from data. Canonicalization and canonical conditions provide a *gauge* (e.g. a canonicalizer Ψ and an associated group element $\kappa(\cdot)$) that allows a non-equivariant model to implement equivariant behavior via an explicit formula.

We assume an exact canonicalizer model, i.e., a canonicalizer provides $\Psi : \mathbb{R}^d \to S$ (slice) and $\kappa : \mathbb{R}^d \to \mathcal{G}$ such that $\Psi(x) = \kappa(x)^{-1} \cdot x$ on a full-measure set (up to stabilizers).

Proposition C.17 (Canonical-condition lifting realizes equivariant functions)**.** *Let $\mathcal{A}$ be any function class on the slice S (not necessarily equivariant). Define the induced ambient class using the canonical condition:*

$$\mathcal{H}_{\mathrm{lift}}(\mathcal{A}) := \big\{\, h(x) = \kappa(x) \cdot a(\Psi(x)) \;:\; a \in \mathcal{A} \,\big\}. \tag{53}$$

Then every $h \in \mathcal{H}_{\mathrm{lift}}(\mathcal{A})$ is $\mathcal{G}$-equivariant (on the set where Ψ, κ are consistent). Moreover, if $\mathcal{A}$ is universal on S, then $\mathcal{H}_{\mathrm{lift}}(\mathcal{A})$ can approximate any equivariant target function on $\mathbb{R}^d$ up to the usual stabilizer caveats.

Proof. Equivariance follows from $\Psi(g \cdot x) = \Psi(x)$ and $\kappa(g \cdot x) = g\kappa(x)$ (in the free-action regime):

$$h(g \cdot x) = \kappa(g \cdot x) \cdot a(\Psi(g \cdot x)) = g\kappa(x) \cdot a(\Psi(x)) = g \cdot h(x). \tag{54}$$

Universality is inherited because any equivariant function is determined by its restriction to the slice. □

Remark C.18 (Why this is a real "shortcut" for non-equivariant networks). A generic non-equivariant network that only sees x must *implicitly* learn the equivariant structure. Providing $(\Psi(x), \kappa(x))$ (or a discrete proxy such as canonical rank/frame ID) allows implementing equivariance by construction as in Proposition C.17. This can reduce approximation complexity and improve optimization.

To summarize, slice training does *not* claim to reduce the ambient Bayes risk equation 48. The population optimum for the same ambient regression problem is unchanged by restriction to the full equivariant function class (Proposition C.15); finite-model training losses also depend on approximation and optimization. Instead, it learns a *different* regression problem on the slice (predicting Δ from $\widetilde{Z}_t$), whose irreducible variance is the first term in Theorem 3.5 and is provably no larger than the ambient one. This is where canonicalization can enable easier training and sampling with fewer steps for non-equivariant models.

C.2.3 Proof for Within-Slice Simplification

This part provides supplementary details on aligned slice priors, OT couplings, and train–test consistency.

Canonicalization eliminates the symmetry-ambiguity term, but the remaining within-slice conditional variance $\mathbb{E}[\mathrm{Var}(\Delta \mid \widetilde{Z}_t)]$ depends on the *slice prior* q_1 and the *slice coupling* $\widetilde{\gamma}$. Section 3.2 explains that after canonicalization, the remaining irreducible term is the *within-slice difficulty* $\mathbb{E}[\mathrm{Var}(\Delta \mid \widetilde{Z}_t)]$, which depends on (i) the slice prior q_1 and (ii) the slice coupling $\widetilde{\gamma}$ (Theorem 3.5). Therefore, we propose two simple yet effective approaches to derive **aligned canonical priors**: (i) *learnable priors* within certain function classes (e.g. a Gaussian with learnable mean and variance); (ii) *KL projections of* q_0 onto certain function classes (e.g. moment-matched Gaussian). For example, among Gaussian canonical priors, the moment-matched Gaussian is $q_1^\star = \mathcal{N}(\mathbb{E}_{q_0}[\widetilde{Z}_0], \mathrm{Cov}_{q_0}(\widetilde{Z}_0))$ (Proposition C.20).

The following collects technical supplements: (i) a closed-form expression showing how *misaligned* Gaussian priors can inflate within-slice variance under product coupling, (ii) why OT/near-Monge couplings are complementary but can be non-unique under symmetry, and (iii) a minimal consistency statement clarifying when training-time OT is compatible with inference and when conditioning causes mismatch.

Gaussian misalignment under product coupling: an exact within-slice variance formula. Canonicalization removes the symmetry-ambiguity term, but even on the slice the conditional variance can remain large if the prior is poorly aligned with the slice geometry. The simplest (and

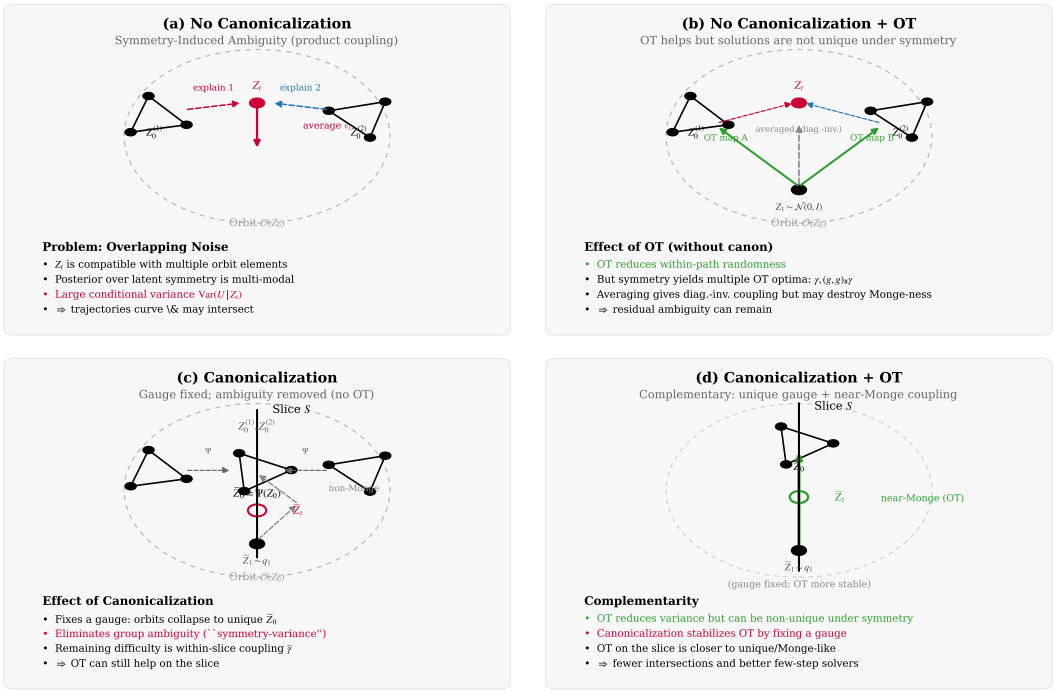


Figure 5: OT and canonicalization act complementarily. OT can reduce conditional variances with or without canonicalization, but OT solutions are generally non-unique in the presence of symmetry (if γ is optimal then $(g,g)_{\#}\gamma$ is also optimal); averaging yields a diagonal-invariant optimum but may destroy Monge-ness. Canonicalization fixes a gauge (collapsing each orbit to a unique slice representative), eliminating group ambiguity and stabilizing OT on the slice.

common) baseline is the product coupling $\widetilde{\gamma} = q_0 \otimes q_1$. The following proposition gives an exact expression for the within-slice conditional covariance in the linear-path (rectified-flow) setting, illustrating how mismatch between Σ_0 and Σ_1 controls the irreducible error.

Proposition C.19 (Closed-form within-slice conditional variance for independent Gaussians)**.** *Assume the slice data and slice prior are Gaussian and independent:*

$$\widetilde{Z}_0 \sim \mathcal{N}(\mu_0, \Sigma_0), \qquad \widetilde{Z}_1 \sim \mathcal{N}(\mu_1, \Sigma_1), \qquad \widetilde{Z}_0 \perp \widetilde{Z}_1.$$

Let $\widetilde{Z}_t = (1-t)\widetilde{Z}_0 + t\widetilde{Z}_1$ *and* $\Delta = \widetilde{Z}_1 - \widetilde{Z}_0$*. For* $0 < t < 1$*, assume* $(1-t)^2\Sigma_0 + t^2\Sigma_1$ *is invertible. Then*

$$\mathrm{Cov}(\Delta \mid \widetilde{Z}_t) = \frac{1}{t^2}\,\mathrm{Cov}(\widetilde{Z}_0 \mid \widetilde{Z}_t), \tag{55}$$

and the conditional covariance admits the closed form

$$\mathrm{Cov}(\widetilde{Z}_0 \mid \widetilde{Z}_t) = \Sigma_0 - (1-t)^2\,\Sigma_0\Big((1-t)^2\Sigma_0 + t^2\Sigma_1\Big)^{-1}\Sigma_0. \tag{56}$$

Consequently, the within-slice irreducible error in flow matching is

$$\mathbb{E}\big[\mathrm{Var}(\Delta \mid \widetilde{Z}_t)\big] = \mathrm{tr}\left(\frac{1}{t^2}\Sigma_0 - \frac{(1-t)^2}{t^2}\,\Sigma_0\Big((1-t)^2\Sigma_0 + t^2\Sigma_1\Big)^{-1}\Sigma_0\right). \tag{57}$$

Proof. The identity $\Delta = (\widetilde{Z}_t - \widetilde{Z}_0)/t$ is algebraic, hence $\mathrm{Cov}(\Delta \mid \widetilde{Z}_t) = \frac{1}{t^2}\mathrm{Cov}(\widetilde{Z}_0 \mid \widetilde{Z}_t)$. Since $(\widetilde{Z}_0, \widetilde{Z}_t)$ is jointly Gaussian with $\mathrm{Cov}(\widetilde{Z}_0, \widetilde{Z}_t) = (1-t)\Sigma_0$ and $\mathrm{Cov}(\widetilde{Z}_t) = (1-t)^2\Sigma_0 + t^2\Sigma_1$, the standard Gaussian conditioning formula yields equation 56. Taking traces gives the final expression. □

Therefore, if one uses an isotropic prior $q_1 = \mathcal{N}(0, I)$ while q_0 is strongly anisotropic (large condition number in Σ_0), then the trace above remains large along directions where Σ_0 dominates I, formalizing

why a *misaligned* simple prior can degrade few-step accuracy even after canonicalization. A practical remedy is to choose an *aligned* prior within a tractable family, e.g. the moment-matched Gaussian $q_1^\star = \mathcal{N}(\mathbb{E}_{q_0}[\widetilde{Z}_0], \mathrm{Cov}_{q_0}(\widetilde{Z}_0))$ (KL projection onto Gaussians), or a learned prior. While it is possible to learn the canonical prior q_1, using KL projections of q_0 is a simple yet effective way. The computation only occurs in pre-processing and does not induce any overhead in training. For instance, among Gaussians, the moment-matched Gaussian is the KL projection of q_0:

Proposition C.20 (Moment-matched Gaussian is the KL-optimal Gaussian approximation)**.** *Let q_0 have a density on $\mathbb{R}^d$, finite differential entropy, finite second moments, and positive-definite covariance. Consider the family $\mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$. The minimizer of $\mathrm{KL}(q_0 \| \mathcal{N}(\mu, \Sigma))$ is*

$$\mu^\star = \mathbb{E}_{q_0}[\widetilde{Z}_0], \qquad \Sigma^\star = \mathrm{Cov}_{q_0}(\widetilde{Z}_0). \tag{58}$$

Proof. This is a standard exponential-family projection: the Gaussian log-density is an affine function of the sufficient statistics $(x, xx^\top)$, so minimizing $\mathrm{KL}(q_0 \| \mathcal{N}(\mu, \Sigma))$ yields moment matching. □

OT/near-Monge couplings: complementary to canonicalization, but symmetry can destroy uniqueness. Given (q_0, q_1), choosing $\widetilde{\gamma}$ close to an OT coupling can make transport more deterministic (Monge-like), shrinking $\mathbb{E}[\mathrm{Var}(\Delta \mid \widetilde{Z}_t)]$ and straightening trajectories (hence benefiting few-step solvers). However, when working in ambient space with symmetric marginals and symmetric costs, OT solutions are generally *not unique*: if γ is optimal then $(g, g)_\# \gamma$ is also optimal. Averaging over g yields a diagonal-invariant optimum, but can destroy Monge-ness by turning a deterministic map into a mixture. Canonicalization fixes a gauge and thus *stabilizes* OT: on the slice, the symmetry-induced degeneracy is reduced, and OT solutions are empirically closer to unique/Monge-like. This is the precise sense in which canonicalization and OT act *complementarily*: canonicalization removes group ambiguity (Theorem 3.5), while OT targets the remaining within-slice difficulty. Figure 5 provides an illustration.

Training-time OT is compatible with inference; mismatch comes from conditioning variables. A frequent confusion is whether using OT in training forces OT at inference. In flow matching / rectified flow, OT changes the *training coupling* $\widetilde{\gamma}$ (and thus the supervision signal), but inference only requires sampling from the *marginal prior* q_1 and integrating the learned dynamics. In particular, if the population vector field is well-defined and sufficiently regular, sampling is *coupling-free* given the correct field: one draws $\widetilde{Z}_1 \sim q_1$ and integrates the learned ODE/SDE to obtain a sample from q_0; no paired endpoints are needed at inference.

The actual train–test mismatch risk arises when the model is conditioned on an auxiliary variable C (e.g. "canonical rank" PE, frame ID, or any deterministic/stochastic output of a canonicalizer). Let $\widetilde{\pi}_1^{\mathrm{tr}}(C, \widetilde{Z}_1)$ be the training-time joint law at the start, and $\widetilde{\pi}_1^{\mathrm{inf}}(C, \widetilde{Z}_1)$ the inference-time one. A sufficient condition for *no mismatch* is:

$$\widetilde{\pi}_1^{\mathrm{tr}}(C, \widetilde{Z}_1) = \widetilde{\pi}_1^{\mathrm{inf}}(C, \widetilde{Z}_1). \tag{59}$$

In particular, if C is a deterministic index inherent to the canonical coordinate system (e.g. the coordinate index i *is* the canonical rank), then this equality holds for any choice of q_1; conversely, if inference computes C as a function of sampled noise in a way that differs from training, mismatch is unavoidable. Finally, note that *canonicalizing noise at inference* (e.g. setting $\widetilde{Z}_1 := \Psi(\varepsilon)$ with $\varepsilon \sim \mathcal{N}(0, I)$) generally induces a complicated slice prior $\Psi_\# \mathcal{N}(0, I)$; unless training uses the same induced prior, this introduces an avoidable prior mismatch and can increase within-slice variance.

D Implementation Details

D.1 Details of Canonicalization Methods

D.1.1 Geometric Spectral Ordering

We define a deterministic canonicalization function $\Psi : \mathcal{M} \to \mathcal{M}$ that maps each molecular graph to a unique representative within its orbit under $S_N \times SE(3)$. The key insight is to leverage the geometric Laplacian constructed from 3D atomic coordinates, whose Fiedler vector encodes the molecular connectivity structure in a continuous, rotation-invariant manner.

Geometric Laplacian construction and canonical permutation. The construction of the geometric Laplacian and the calculation of the (signed) Fiedler vector can be cleanly summarized as the pseudo-algorithm in Algorithm 1.

Algorithm 1 Geometric Laplacian and signed Fiedler vector

Input: coordinates $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbb{R}^{N\times 3}$, bond set E.
Compute mean bond length $\bar{d}_{\text{bond}} \leftarrow \frac{1}{|E|}\sum_{(i,j)\in E} \|\mathbf{x}_i - \mathbf{x}_j\|$.
Set bandwidth $\sigma^2 \leftarrow 4\,\bar{d}^2_{\text{bond}}$.
for $i = 1, \ldots, N$ **do**
  for $j = 1, \ldots, N$ **do**
    if $i = j$ **then**
      $W_{ij} \leftarrow 0$.
    else
      $W_{ij} \leftarrow \exp\big(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2)\big)$.
    end if
  end for
end for
Degree matrix $\mathbf{D} \leftarrow \mathrm{diag}(\sum_j W_{ij})$.
Random-walk Laplacian $L_{\text{rw}} \leftarrow \mathbf{D}^{-1}(\mathbf{D} - \mathbf{W})$.
Compute eigenpairs of L_{rw}: $0 = \lambda_1 \le \lambda_2 \le \cdots$ with eigenvectors $\{u_k\}$.
Fiedler vector $u_2 \leftarrow$ eigenvector of λ_2.
Centroid $\bar{\mathbf{x}} \leftarrow \frac{1}{N}\sum_i \mathbf{x}_i$ and mean radius $\bar{d} \leftarrow \frac{1}{N}\sum_i \|\mathbf{x}_i - \bar{\mathbf{x}}\|$.
Fix sign by centroid-direction convention:
  $u_2 \leftarrow \mathrm{sign}\big(\sum_i u_{2,i}(\|\mathbf{x}_i - \bar{\mathbf{x}}\| - \bar{d})\big) \cdot u_2$.
Output: signed Fiedler vector u_2.

Algorithm 2 Spectral $SO(3)$ Canonicalization

Require: Atomic coordinates $\mathbf{X} \in \mathbb{R}^{N\times 3}$ and signed Fiedler vector v (computed by Algorithm 1)
Ensure: Canonicalized coordinates $\mathbf{X}'$

1: Apply spectral ordering $\pi^* = \mathrm{argsort}(v)$ using the signed vector returned by Algorithm 1
2: Let head index h be rank 0 and tail index t be rank $N-1$
3: Choose anchor index

$$a = \mathrm{argmax}_{k\in[\lfloor N/3\rfloor, \lfloor 2N/3\rfloor)} \big\|(\mathbf{x}_k - \mathbf{x}_h) \times (\mathbf{x}_t - \mathbf{x}_h)\big\|$$

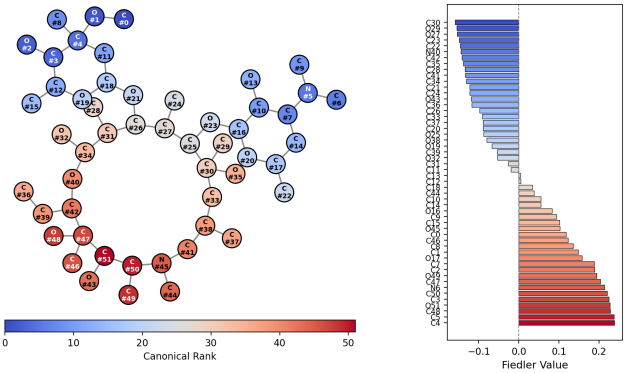
4: Compute longitudinal axis $\mathbf{e}_1 \leftarrow \frac{\mathbf{x}_t - \mathbf{x}_h}{\|\mathbf{x}_t - \mathbf{x}_h\|}$
5: Compute plane normal $\mathbf{n} \leftarrow \mathbf{e}_1 \times (\mathbf{x}_a - \mathbf{x}_h)$
6: Normalize normal $\mathbf{e}_3 \leftarrow \frac{\mathbf{n}}{\|\mathbf{n}\|}$
7: Complete right-handed basis $\mathbf{e}_2 \leftarrow \mathbf{e}_3 \times \mathbf{e}_1$
8: Assemble rotation matrix $\mathbf{R} \leftarrow [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]^\top$
9: Center and rotate $\mathbf{X}' \leftarrow (\mathbf{X} - \bar{\mathbf{X}})\mathbf{R}^\top$

The canonical ordering is then: $\pi^* = \mathrm{argsort}(u_2)$ and the canonicalized molecule is $\Psi^{(\mathrm{perm})}(\mathbf{Z}) = \pi^*(\mathbf{Z})$. The Fiedler vector approximately captures the graph Laplacian's fundamental mode of oscillation, providing a "core-to-periphery" ordering. Atoms with similar Fiedler values tend to be spatially close and structurally equivalent, which induces a natural generation order from molecular core to functional groups; see Figure 6 for illustrative examples.

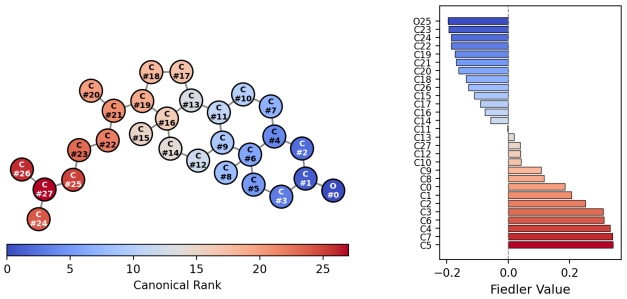
Canonical $SO(3)$ frame. The above procedure determines a canonical order on S_N, based on which we further (optionally) process the $SO(3)$ symmetry using Algorithm 2.

D.1.2 Alternative Orderings

Graph canonicalization is a widely studied topic and there are also other existing canonicalization methods. However, most of them are defined for permutation in abstract graphs (Zhao et al., 2024; Ma et al., 2023a; Dong et al., 2024) instead of geometric graphs.



(a) Azithromycin (macrocycle): the spectral ordering is core-centric, concentrating the macrocycle near the spectral center and mapping peripheral sugar moieties toward the sequence terminals.



(b) Cholesterol (anisotropic): the spectral ordering unfolds along the principal axis from the polar head through the rigid tetracyclic core to the flexible hydrocarbon tail, while preserving local connectivity.

Figure 6: Spectral canonicalization via the (signed) Fiedler vector produces stable, locality-preserving atom orderings for two structurally different molecules (blue: start, red: end), illustrating orderings across distinct molecular geometries.

We also implement two simpler orderings for ablation:

1. Structural Multihop (inspired by PARD (Zhao et al., 2024)): Iterative degree peeling using weighted multihop degrees $w_K(v) = \sum_{k=1}^{K} d_k(v) \cdot N^{K-k}$, where $d_k(v)$ counts nodes at exactly k hops.
2. Atomic Numbering: Priority-based ordering by atomic number (e.g., heavy/rare atoms first, hydrogen last).

However, we experimentally find that our structure-aware geometric spectral ordering outperforms all these methods (Table 14), validating its effectiveness.

D.2 Canonical Diffusion and CanonFlow

We present additional algorithmic and architectural details for canonical diffusion and flow matching.

D.2.1 Details of canonical diffusion and flow matching

We have already described the overall framework of canonical diffusion or flow matching in Section 4.2. We now provide additional details in this part.

Canonical positional encoding. To explicitly break permutation equivariance, we inject positional information derived from the canonical rank into the model architecture. For each atom i with

normalized rank $r_i = \text{rank}_i/N \in [0, 1)$, we compute a sinusoidal positional encoding:

$$\text{PE}(r, 2k) = \sin\left(\frac{r \cdot M}{10000^{2k/d_{\text{pe}}}}\right), \qquad \text{PE}(r, 2k+1) = \cos\left(\frac{r \cdot M}{10000^{2k/d_{\text{pe}}}}\right) \tag{60}$$

where $M = 10000$ is the maximum position scale and d_{pe} is the encoding dimension. This encoding is concatenated with atom features before the initial projection. This completely breaks permutation equivariance at the architecture level, allowing the model to distinguish atoms by their structural role. Note that this PE can be applied to either our novel Canon architecture, or other existing architectures to make a (non-equivariant) canonicality-aware counterpart. The *PE-drop* mechanism replaces $\mathbf{R}$ with a learned "fake PE" embedding with some probability p_{drop} during training, which is used for classifier-free guidance w.r.t. canonical conditions.

Adaptive position-dependent prior. For categorical features, our baseline uses a uniform prior q_1. We replace this with a position-dependent prior that encodes empirical correlations between canonical position and atom type:

$$q_1^{\text{eff}}(c \mid r) = \beta_r \cdot p_{\text{prior}}(c) + (1 - \beta_r) \cdot p_r(c \mid r) \tag{61}$$

where $r = r_i$ is the normalized rank defined above, β_r is a mixing coefficient (typically 0.1 for categorical features), p_{prior} is the prior distribution such as p_{uniform}, and $p_r(\cdot|r)$ is the empirical distribution estimated from the training set.

For *prior estimation*, we discretize $r \in [0, 1)$ into K bins and compute:

$$P(c \mid B_k) = \frac{\text{Count}(c, k) + \epsilon}{\sum_{c'}(\text{Count}(c', k) + \epsilon)} \tag{62}$$

for categorical data; we use a Gaussian with sufficient statistics mean and variance in the dataset for continuous coordinates. At runtime, we use linear interpolation between adjacent bins for continuous r:

$$p_r(c|r) = (1 - \delta) \cdot P(c|B_{k_0}) + \delta \cdot P(c|B_{k_1}) \tag{63}$$

where $k_0 = \lfloor r \cdot K \rfloor$, $k_1 = \min(k_0 + 1, K - 1)$, and $\delta = r \cdot K - k_0$.

D.2.2 Canon Architecture

The Semla architecture proposed by Irwin et al. (2024) maintains two hidden states corresponding to $\mathbf{X}, \mathbf{H}$ in each layer. In our *Canon* architecture, we additionally incorporate a third hidden state to update and refine canonical information in each layer by interacting with other atom and bond features. Figure 7 summarizes the Canon architecture.

Input processing. We implement a three-stream molecular dynamics network, **Canon**, which augments a Semla-like equivariant message passing trunk with an explicit *canonical-rank stream* and optional canonical positional encodings (PE). Concretely, at diffusion/flow time t, the model takes as input

$$\mathbf{X}_t \in \mathbb{R}^{B\times N\times 3}, \qquad \mathbf{H}_t \in \mathbb{R}^{B\times N\times d_h}, \qquad \mathbf{R} \in \mathbb{R}^{B\times N},$$

where $\mathbf{X}_t$ are coordinates, $\mathbf{H}_t$ are node-wise invariant features (e.g. atom features, optional conditioning), and $\mathbf{R}$ is the canonical rank (or its normalized variant). The positional encoding processing is identical to the general description above. Optional self-conditioning provides a previous estimate $\hat{\mathbf{X}}_0$, and optionally a previous rank estimate $\hat{\mathbf{R}}$. In the implementation, self-conditioning is realized by stacking $(\mathbf{X}_t, \hat{\mathbf{X}}_0)$ and projecting into K coordinate sets via a linear map Π_{CS}:

$$\mathbf{CS}_t \;=\; \Pi_{CS}\big([\mathbf{X}_t, \hat{\mathbf{X}}_0]\big) \in \mathbb{R}^{B\times K\times N\times 3}, \tag{64}$$

with masking applied on padded atoms.

We append a size embedding $\text{Emb}(|V|)$ and optional canonical PE to the invariant features and project them into the trunk width d:

$$\widetilde{\mathbf{H}}_t \;=\; \phi_{\text{in}}\Big([\mathbf{H}_t, t, \text{Emb}(|V|), \text{PE}(\mathbf{R})]\Big) \in \mathbb{R}^{B\times N\times d}, \tag{65}$$

where ϕ_{in} is a two-layer MLP with SiLU nonlinearity.

CanonDynamics: message-passing layers with node/coord/rank streams. The trunk (`CanonDynamics`) stacks L copies of a message passing layer, optionally with edge features. Each layer updates three coupled states:

$$(\mathbf{CS}, \mathbf{H}, \mathbf{R}) \mapsto (\mathbf{CS}', \mathbf{H}', \mathbf{R}'),$$

In each *edge message layer*, we construct pairwise messages using *(i)* projected node features, *(ii)* projected rank features, and *(iii)* geometric features derived from coord-set dot-products. Specifically, after normalization we compute

$$\mathbf{G}_{ij}^{(k)} \;=\; \langle \mathbf{CS}_i^{(k)},\, \mathbf{CS}_j^{(k)} \rangle, \qquad k = 1, \ldots, K, \tag{66}$$

and concatenate

$$\boldsymbol{m}_{ij} \;=\; \mathrm{MLP}\Big([\mathbf{W}_h \boldsymbol{h}_i,\, \mathbf{W}_h \boldsymbol{h}_j,\, \mathbf{W}_r \boldsymbol{r}_i,\, \mathbf{W}_r \boldsymbol{r}_j,\, \mathbf{G}_{ij}^{(1:K)},\, \boldsymbol{e}_{ij}] \Big), \tag{67}$$

where $\boldsymbol{e}_{ij}$ is an optional edge feature input, and $\mathbf{W}_h, \mathbf{W}_r$ are learnable projection weights.

Next we adopt attention-based updates for node/coord/rank features. The message tensor is split into three channels (plus optional edge-out):

$$\boldsymbol{m}_{ij} \;=\; \big(\boldsymbol{m}_{ij}^{\mathrm{node}},\, \boldsymbol{m}_{ij}^{\mathrm{coord}},\, \boldsymbol{m}_{ij}^{\mathrm{rank}},\, \boldsymbol{m}_{ij}^{\mathrm{edge}}\big), \tag{68}$$

which drive three attention modules:

$$\begin{aligned} \mathbf{H} &\leftarrow \mathbf{H} + \mathrm{Attn}_{\mathrm{node}}(\mathbf{H}, \boldsymbol{m}^{\mathrm{node}}), \\ \mathbf{CS} &\leftarrow \mathbf{CS} + \mathrm{Attn}_{\mathrm{coord}}(\mathbf{CS}, \boldsymbol{m}^{\mathrm{coord}}), \\ \mathbf{R} &\leftarrow \mathbf{R} + \mathrm{Attn}_{\mathrm{rank}}(\mathbf{R}, \boldsymbol{m}^{\mathrm{rank}}). \end{aligned} \tag{69}$$

interleaved with feed-forward residual blocks consisting of equivariant or invariant MLPs for $(\mathbf{H}, \mathbf{CS})$ and $\mathbf{R}$.

Output heads. After L layers, Canon produces coordinate updates through a normalization + linear head:

$$\widehat{\mathbf{X}} \;=\; \mathrm{Head}_X\big(\mathrm{Norm}(\mathbf{CS})\big) \in \mathbb{R}^{B \times N \times 3}, \tag{70}$$

and predicts atom type and charge logits via linear classifiers on the final node features. If enabled, it also predicts a rank score per node via a linear head and min–max normalization to $[0, 1]$. Optional edge features are refined by a bond-refinement module and projected to edge-type logits.

D.3 Training

Rank noise for robustness. To reduce the train-inference gap (since ground-truth ranks are unavailable during sampling), we add noise to canonical ranks during training:

$$r_{\text{noised}} = r + \mathcal{N}(0, \sigma_r^2) \cdot (t) \tag{71}$$

The noise magnitude decreases as t approaches zero to avoid large perturbations when the molecule is nearly reconstructed.

Auxiliary rank prediction. Optionally, the model predicts canonical rank as an auxiliary task:

$$\hat{r} = \sigma\big(\mathrm{MLP}(h_i^{(L)})\big) \in [0, 1] \tag{72}$$

with min-max normalization across atoms and training loss:

$$\mathcal{L}_{\text{rank}} = \lambda_r \cdot \mathbb{E}\left[\|r^* - \hat{r}\|^2\right] \tag{73}$$

This enables dynamic rank estimation during inference.

Optimal transport annealing. We observe that optimal transport (OT) matching between noise and target molecules reduces flow variance but may limit generalization. We implement **OT annealing**: the probability of using OT decreases linearly during training:

$$p_{\text{OT}}(\text{epoch}) = \max\left(0, 1 - \frac{\text{epoch}}{\text{max epochs}}\right) \tag{74}$$

This allows the model to benefit from OT alignment early in training while learning to handle arbitrary noise-target pairings later.

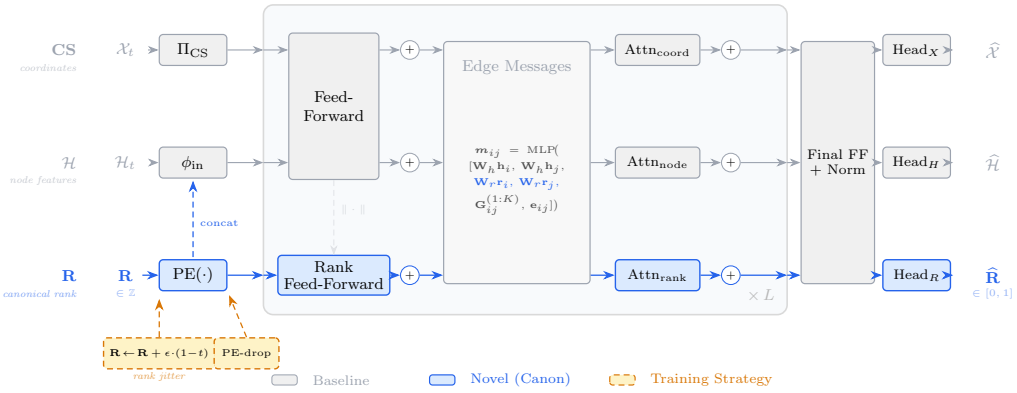


Figure 7: Overview of the Canon architecture. Three parallel streams—coordinates (**CS**), node features ($\mathcal{H}$), and canonical rank (**R**)—are updated through L message-passing layers consisting of feed-forward blocks, pairwise edge messages, and attention-based aggregation. Gray blocks denote components inherited from Semla; blue blocks denote novel canonical-rank components introduced by Canon. Rank information is jittered prior to positional encoding and dropped afterward as a training strategy.

Algorithm 3 Gap-free canonical training with optional OT coupling (no OT needed at inference)

Require: Dataset (G, X); symmetry group $G \in \{S_N,\ S_N \times SO(3)\}$; canonicalizer Ψ_G producing rank-ordered canonical coordinates $\widetilde{X}$.

1: **Group setup:** choose Ψ_G and the canonical-space prior q_1 accordingly.
2: Canonicalize each training example once: $\widetilde{Z}_0 = \text{vec}(\Psi(X))$; treat rank as the fixed coordinate index.
3: Choose a marginal prior q_1 in canonical space (Aligned or Isotropic).
4: **for** each training iteration **do**
5: Sample minibatch $\{\widetilde{z}_0^i\} \sim q_0$ and noise minibatch $\{\widetilde{z}_1^j\} \sim q_1$.
6: Construct a coupling:
- Product: pair $(\widetilde{z}_0^i, \widetilde{z}_1^i)$; or
- OT/Sinkhorn: compute plan Π minimizing $\sum_{i,j} \Pi_{ij} \|\widetilde{z}_0^i - \widetilde{z}_1^j\|^2$, then sample pairs from Π.

7: Train the flow/diffusion model on canonical coordinates using fixed/predicted rank embeddings $\text{PE}(i)$.
8: **end for**

Training objective. The overall training loss combines coordinate regression, categorical cross-entropy for discrete features, and optional rank prediction:

$$\mathcal{L}_{\text{coord}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{\mathcal{X}}_{0,i} - \mathcal{X}_{0,i}\|^2, \qquad \mathcal{L}_{\text{type}} = \frac{1}{N} \sum_{i=1}^{N} \text{CE}(\hat{\mathbf{h}}_{0,i}, \mathbf{h}_{0,i}) \tag{75}$$

$$\mathcal{L}_{\text{bond}} = \frac{1}{N^2} \sum_{i,j} \text{CE}(\hat{\mathbf{E}}_{0,ij}, \mathbf{E}_{0,ij}), \qquad \mathcal{L}_{\text{rank}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{r}_i - r_i^*)^2 \tag{76}$$

$$\mathcal{L} = \mathcal{L}_{\text{coord}} + \lambda_{\text{type}} \mathcal{L}_{\text{type}} + \lambda_{\text{bond}} \mathcal{L}_{\text{bond}} + \lambda_{\text{charge}} \mathcal{L}_{\text{charge}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}} \tag{77}$$

where default weights are $\lambda_{\text{type}} = 0.2$, $\lambda_{\text{bond}} = \lambda_{\text{charge}} = 1.0$ and $\lambda_{\text{rank}} = 0.1$. Following Irwin et al. (2024), we set $\lambda_{\text{bond}} = 0.5$ for the QM9 dataset.

The complete training procedure can be summarized in Algorithm 3.

D.4 Sampling

Algorithm 4 Unified few-step canonical sampling (Regime A/B + options)

Require: Symmetry group $G \in \{S_N,\ S_N \times SO(3)\}$; trained model v_θ; baseline prior q_1; step times $1 = t_K > \cdots > t_0 = 0$.
Require: Sampling regime flag $\mathsf{Regime} \in \{\mathrm{A}, \mathrm{B}\}$.
Require: Options: (i) aligned prior; (ii) PCS (if enabled, apply Ψ after each step; requires $\mathsf{Regime} = \mathrm{B}$).
Note (rank conditioning): if rank is used as a condition/PE, treat it as the *fixed coordinate index* and *do not* recompute it from noise.
(Optional prior choice): optionally replace q_1 with an alternative prior.
Sample $\widetilde{Z}_{t_K} \sim q_1$ in the fixed canonical coordinate system.
for $k = K, \ldots, 1$ **do**
  $\widetilde{Z}_{t_{k-1}} \leftarrow \mathrm{Step}(\widetilde{Z}_{t_k}, v_\theta, t_k \to t_{k-1})$ {ODE/SDE solver}
  if $\mathsf{Regime} = \mathrm{B}$ **then**
    $\widetilde{Z}_{t_{k-1}} \leftarrow \Psi(\widetilde{Z}_{t_{k-1}})$
  else
    Note: do *not* re-canonicalize / re-rank / project during sampling in Regime A; keep rank indices fixed.
  end if
end for
Output $\widetilde{Z}_{t_0}$.
Optional invariance restoration: sample $g \sim \lambda(G)$ and output $g \cdot \widetilde{Z}_{t_0}$.

Recovering invariance through Haar randomization. The critical observation is that canonical samples, while not themselves drawn from an invariant distribution, can be *lifted* to an invariant distribution through a simple post-processing step. As an immediate result of Lemma 3.1, the following proposition, which generalizes the observation in Yan et al. (2023) to arbitrary compact groups, provides the theoretical guarantee in sampling invariant distributions:

Proposition D.1 (Post-hoc randomization yields invariant sampling)**.** *Let $\tilde{\mu}$ be any distribution on $\mathcal{M}$. Define the randomized distribution*

$$\mu := \int_{\mathcal{G}} (g\cdot)\#\tilde{\mu}\,\mathrm{d}\lambda(g). \tag{78}$$

Then μ is $\mathcal{G}$-invariant. Moreover, sampling $g \sim \lambda$ and $\tilde{\mathbf{Z}} \sim \tilde{\mu}$ and outputting $g \cdot \tilde{\mathbf{Z}}$ produces μ.

Proof. For any $h \in \mathcal{G}$,

$$(h\cdot)\#\mu = \int_{\mathcal{G}} (hg\cdot)\#\tilde{\mu}\,\mathrm{d}\lambda(g) = \int_{\mathcal{G}} (g'\cdot)\#\tilde{\mu}\,\mathrm{d}\lambda(g') = \mu, \tag{79}$$

where we changed variables $g' = hg$ and used left-invariance of Haar. □

Remark D.2 (Canonicalization vs. randomization). Randomization alone enforces invariance but does not simplify the learning problem. Canonicalization enforces a *gauge choice* that can substantially reduce multimodality/variance induced by symmetry and thereby accelerate training convergence, while randomization then restores invariance at the end. The combination of canonicalized training with post-hoc randomization thus achieves the best of both worlds: efficient learning and provably invariant generation.

Dynamic rank estimation. Since sampling can be done using a fixed canonical rank as the condition, dynamic rank updates are completely optional.

A practical few-step stabilizer is to project intermediate states back to the slice (Regime B in Algorithm 4). In particular, to keep rank conditions and partially noisy samples consistent, we allow projection-to-slice sampling, which we term Projected Canonical Sampling (PCS) as detailed in Algorithm 5. Note that projecting the data to the canonical slice while keeping the PE ranks fixed is conceptually equivalent to keeping the data in the original order but updating the canonical rank to align the two.

We support two strategies to estimate canonical rank in the absence of ground truth in Algorithm 5:

1. Predict Mode: Use the model's rank prediction output $\hat{r}$ directly for warping.
2. Canonicalize Mode: Periodically recompute canonical rank from intermediate predictions: Discretize current atom type predictions: $\hat{a}_i = \text{argmax}(\text{atomics}_i)$. Then apply the same canonicalization algorithm to the predicted molecule. Update rank every K steps when $t \geq T$.

We experimentally find that the predict mode is always better, likely due to the instability of canonicalization algorithms for intermediate noisy states; instead, self-prediction integrates this procedure into the model learning, facilitating robustness and expressivity.

Remarkably, if the dynamic rank estimation is disabled (Regime A in Algorithm 4), the model always conducts a "conditional generation" task given the canonical rank conditions, which is also a valid task with no train-test gap.

The complete sampling procedure can be summarized in Algorithm 4.

Algorithm 5 Projected Canonical Sampling (PCS)

Require: Canonicalizer Ψ_ϕ; learned slice model (score or vector field) m_θ; time grid $1 = t_K > \cdots > t_0 = 0$.

1: Initialize $\tilde{\mathbf{Z}}_{t_K} \sim q_1$.
2: **for** $k = K, \ldots, 1$ **do**
3: Take one reverse step on the slice: $\hat{\tilde{\mathbf{Z}}} \leftarrow \text{Step}(\tilde{\mathbf{Z}}_{t_k}, m_\theta, t_k \to t_{k-1})$.
4: Project: $\tilde{\mathbf{Z}}_{t_{k-1}} \leftarrow \Psi_\phi(\hat{\tilde{\mathbf{Z}}})$.
5: **end for**
6: Output canonical sample $\hat{\mathbf{Z}}_0 = \tilde{\mathbf{Z}}_{t_0}$ and optionally randomize by $g \sim \lambda$.

E Experimental Details and Additional Results

E.1 Experimental Details

Baseline references. The baselines in Tables 1 and 2 are EDM (Hoogeboom et al., 2022), GCDM (Morehead & Cheng, 2024), MUDiff (Hua et al., 2024), FlowMol (Dunn & Koes, 2024), MiDi (Vignac et al., 2023), EQGAT-diff (Le et al., 2023), and SemlaFlow (Irwin et al., 2024).

E.1.1 Evaluation metrics

We evaluate generated molecules using the following metrics. Let $\mathcal{M} = \{m_i\}_{i=1}^N$ denote the set of generated molecules, and let $\tilde{m}_i$ denote the RDKit force-field optimized conformation of m_i.

- **Validity.** The fraction of generated molecules that pass RDKit chemical validity checks.
- **Atom Stability.** The proportion of all generated atoms whose explicit valence lies within a predefined allowed range.
- **Molecule Stability.** The fraction of generated molecules in which *every* atom is stable.
- **Uniqueness.** Among molecules that can be converted to canonical SMILES, the fraction of distinct SMILES strings.
- **Novelty.** Among molecules with valid canonical SMILES, the fraction absent from the training set.
- **Opt-RMSD.** The average root-mean-square deviation between generated and optimized conformations:

$$\text{Opt-RMSD} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \text{RMSD}(m_i, \tilde{m}_i),$$

where $\mathcal{I} = \{i : \text{optimization succeeds and RMSD is finite}\}$. This quantifies how far the generated 3D geometry deviates from a local energy minimum.

- **NFE.** The Number of Function Evaluations, i.e. the total number of neural network forward passes required by the ODE solver during sampling. NFE serves as a hardware-agnostic measure of generation cost.
- **Sampling Time.** The wall-clock time to generate the full evaluation set (N=1,000 molecules).

E.1.2 HYPER-PARAMETERS.

Most of our hyperparameters follow Irwin et al. (2024) for both Canonical SemlaFlow and our CanonFlow. All of our backbones consist of L=12 equivariant message-passing layers (d_{model}=384, d_{msg}=128, d_{edge}=128, H=32 attention heads, S=64 coordinate sets with per-set length normalization), with two edge-aware layers at the input and output. A learnable molecule-size embedding of dimension 64 is concatenated to each atom's input features. All output heads (atom type, charge, bond type, and the optional canonical-rank head) are two-layer MLPs with SiLU activation. Self-conditioning is employed by default: during training, with probability 0.5 the model first produces a preliminary prediction with zeroed conditioning inputs, which is then fed back as additional context; during inference it is applied at every integration step.

Models are trained for 200 epochs on GEOM-DRUG and 300 epochs on QM9 with Adam (lr=3×10^{-4}, `amsgrad`, no weight decay), a linear warm-up (10 000 steps for GEOM-DRUG, 2 000 for QM9) followed by a constant learning rate, gradient clipping at global norm 1.0, and bucket-based dynamic batching with a cost budget of 4096 and linear cost scaling. An exponential moving average (EMA, decay 0.999) of model parameters is maintained and used for all inference. All training is conducted in FP32 precision. The total training time is approximately 10 hours for QM9 and 28 hours for GEOM-DRUG on a single NVIDIA A100 80GB GPU.

Coordinates are interpolated linearly, $\mathbf{x}_t = (1-t)\,\mathbf{x}_0 + t\,\mathbf{x}_1$, with additive Gaussian noise (σ=0.2). Atom types and bond types use the uniform-sample categorical interpolation Campbell et al. (2022). The interpolation time is drawn from $\mathrm{Beta}(2,1)$. Equivariant optimal transport (Kabsch alignment + Hungarian matching) is optionally applied to coordinate pairs. When classifier-free guidance (CFG) is enabled, the positional-encoding dropout rate defaults to p_{drop}=0.1.

E.1.3 COMPUTATIONAL COST

One-time preprocessing. Canonicalization timings are measured on an AMD EPYC 9654 CPU using one thread, five molecules per atom count, and the median of seven repetitions. At $N = 70$ atoms, permutation ordering takes 0.637 ms per molecule, while frame construction takes 1.511 ms, for a total of 2.148 ms. Extrapolated dataset-level preprocessing costs are below 0.7% of the corresponding representative training times (Table 6). These estimates exclude data I/O and cache serialization.

Table 6: Estimated preprocessing cost relative to representative single-GPU training runs.

Dataset	Training molecules	Preprocessing (s)	Training (h)	Ratio (%)
QM9	96,663	199.7	9	0.62
GEOM-DRUG	240,990	515.7	40	0.36

Training-time comparison. In a separate GEOM-DRUG comparison, training takes 60.8 h for SemlaFlow, 61.3 h for canonicalized SemlaFlow, and 61.7 h for CanonFlow. Another same-device comparison records 10 h 13 min for SemlaFlow and 10 h 15 min for canonicalized SemlaFlow. These measurements indicate negligible computational overhead in wall-clock time, proving the efficiency of our canonicalization method.

E.2 ADDITIONAL RESULTS

E.2.1 NUMERICAL VALIDATION OF THEORETICAL RESULTS

Stability of molecular canonicalization. We now report spectral degeneracy and the free-action regime in realistic molecules. Symmetric graph substructures do not imply a non-trivial stabilizer of

Table 7: Spectral diagnostics of practical molecules. Full test splits are used for eigengaps and exact ties.

Diagnostic	QM9	GEOM-DRUG
Median / first-percentile γ	$0.224/0.023$	$0.469/0.064$
$\gamma < 10^{-3}$	$5/12{,}929$ (0.039%)	$1/10{,}000$ (0.010%)
Exact ties in sorted Fiedler components	$1/12{,}929$ (0.008%)	$3/10{,}000$ (0.030%)
MAP fallback	$0/4{,}000$	$0/3{,}000$
$SO(3)$ frame-degeneracy flag	$0/4{,}000$	$0/3{,}000$

Table 8: Structural perturbation audit of the complete canonical pose. RMSD-ratio quantiles are conditional on preserved ordering and do not describe pairs that cross a rank boundary.

Dataset	Noise σ (Å)	Order preserved (%)	Median RMSD ratio	95th percentile
QM9	10^{-4}	91.5	1.029	1.75
QM9	10^{-3}	84.6	1.027	1.66
GEOM-DRUG	10^{-4}	98.8	1.006	1.28
GEOM-DRUG	10^{-3}	85.8	1.006	1.26

the complete three-dimensional conformer. We examine the relative eigengap

$$\gamma = \frac{|\lambda_3 - \lambda_2|}{\max(|\lambda_2|, |\lambda_3|)}. \tag{80}$$

The median gaps are 0.224 on QM9 and 0.469 on GEOM-DRUG (Table 7). Gaps below 10^{-3} occur in 0.039% and 0.010% of the respective test splits. This threshold triggers an expanded retained eigenspace with MAP applied within it; it does not itself indicate canonicalization failure. Separately audited subsets contain no MAP fallback or frame-degeneracy flags. These diagnostics support the practical relevance of Assumption 2.5.

To quantify stability under structural perturbations, we add isotropic coordinate noise to 1,499 QM9 and 800 GEOM-DRUG molecules and independently canonicalize each original and perturbed geometry using the full $S_N \times SO(3)$ map. Table 8 reports order preservation and the ratio of output RMSD to input RMSD. Conditional on preserved order, median ratios lie between 1.006 and 1.029, indicating little typical amplification. Order preservation decreases at larger noise levels, and rank crossings can produce discrete jumps. The evidence therefore supports piecewise local stability.

Empirical analysis of symmetry-induced variance. We first empirically verify the variance decomposition result indicated in Theorem 3.5 using practical datasets. Let U denote the flow-matching regression target, Z_t the interpolated state, and g the latent group action associated with a symmetry group G. For vector-valued targets, variance denotes the trace of the conditional covariance. With conditioning on t implicit, the law of total variance gives

$$V_{\text{amb}}(G) = \mathbb{E}_{t,Z_t}[\text{Var}(\mathbb{E}[U \mid Z_t, g] \mid Z_t)], \tag{81}$$

$$V_{\text{within}}(G) = \mathbb{E}_{t,Z_t}[\mathbb{E}[\text{Var}(U \mid Z_t, g) \mid Z_t]], \tag{82}$$

$$V_{\text{ambient}} = \mathbb{E}_{t,Z_t}[\text{Var}(U \mid Z_t)] = V_{\text{amb}}(G) + V_{\text{within}}(G). \tag{83}$$

Table 9 shows that permutation ambiguity accounts for 37.3% of the ambient variance on QM9 and 38.9% on GEOM-DRUG. Including rotations increases the ambiguity component by 0.52 and 0.34, respectively, corresponding to a further 4.5% and 2.5% of the ambient variance. Permutation ambiguity therefore accounts for most of the measured symmetry contribution in these settings. The smaller additional rotational component motivates evaluating S_N and $S_N \times SO(3)$ separately, and we also attribute the slightly lower performance of $S_N \times SO(3)$ to its instability when determining the axis. Therefore, we believe S_N alone is sufficient and could largely improve empirical performance.

We also measure the training losses of coordinate regression w/ and w/o canonicalization, which measure achieved regression errors rather than the irreducible Bayes risk. Canonical training reduces the EMA coordinate loss from 0.165 to 0.104 on QM9 and from 0.211 to 0.095 on GEOM-DRUG. These observations are consistent with reduced regression difficulty.

Table 9: Variance decomposition and coordinate training loss with exponential moving averages (EMA). The two variance components sum to the ambient value for each group.

		S_N		$S_N \times SO(3)$		EMA coordinate loss	
Dataset	Ambient	Ambiguity	Within	Ambiguity	Within	Baseline	Canonical
QM9	11.67	4.35	7.32	4.87	6.80	0.165	0.104
GEOM-DRUG	13.55	5.27	8.28	5.61	7.94	0.211	0.095

E.2.2 Ablation Studies

In addition to our state-of-the-art main results presented in Section 5, we provide more comprehensive ablation studies here.

Superior few-step generation with CanonFlow. Table 10 presents few-step generation results of our CanonFlow (trained with OT annealing), which consistently outperforms the SemlaFlow baseline w/o canonicalization, and the advantage is still significant even with only 50 steps. This further supports our claim that canonicalization not only improves the upper bound of generative models, but also provides a strong guidance signal through the conditions to facilitate few-step generation.

Table 10: CanonFlow w/ our novel Canon architecture on GEOM-DRUG. Canonicalization is conducted on S_N.

Model	Mol Stab ↑	Valid ↑	NFE
EQGAT-diff	$93.4_{\pm 0.21}$	$94.6_{\pm 0.24}$	500
SemlaFlow_{50} w/ OT	$97.0_{\pm 0.21}$	$93.9_{\pm 0.12}$	50
CanonFlow_{50}	$\mathbf{98.0}_{\pm 0.08}$	$\mathbf{95.4}_{\pm 0.09}$	50
SemlaFlow_{100} w/ OT	$97.3_{\pm 0.08}$	$93.9_{\pm 0.19}$	100
CanonFlow_{100}	$\mathbf{98.4}_{\pm 0.02}$	$\mathbf{95.9}_{\pm 0.08}$	100
Data	100.0	100.0	–

Effects of classifier-free guidance. We further ablate the sampling quality with CFG. To comprehensively evaluate the effects of canonicalized symmetry groups, we report the performance of canonical SemlaFlow on both S_N (trained w/ OT annealing) in Table 12 and $S_N \times SO(3)$ (trained w/o OT) in Table 11. Within an appropriate range, a larger CFG scale tends to improve the validity metric, at the cost of slightly decreasing the molecule stability. Interestingly, the improvement is more significant on the $S_N \times SO(3)$ canonicalization, since the prediction relies more on the canonical information, which benefits from extrapolating the conditional shortcut and unconditional baseline.

Table 11: Canonical SemlaFlow on the GEOM-DRUG dataset w/ classifier-free guidance (CFG). Canonicalization is conducted on $S_N \times SO(3)$.

Model	CFG scale	Mol Stab ↑	Valid ↑	NFE
EQGAT-diff	–	$93.4_{\pm 0.21}$	$94.6_{\pm 0.24}$	500
SemlaFlow_{100}	–	$97.3_{\pm 0.08}$	$93.9_{\pm 0.19}$	100
Canon. SemlaFlow_{100}	1.0	$\mathbf{97.9}_{\pm 0.09}$	$93.8_{\pm 0.30}$	100
Canon. SemlaFlow_{100}	1.5	$97.7_{\pm 0.06}$	$94.2_{\pm 0.23}$	100
Canon. SemlaFlow_{100}	2.0	$97.6_{\pm 0.09}$	$\mathbf{94.4}_{\pm 0.05}$	100
Data	–	100.0	100.0	–

Ablation on aligned prior and OT. In Table 13 we study the interaction between aligned priors and OT for both equivariant and non-equivariant backbones under S_N canonicalization. Without PE, the

Table 12: Canonical SemlaFlow on the GEOM-DRUG dataset w/ classifier-free guidance (CFG). Canonicalization is conducted on S_N.

Model	CFG scale	Mol Stab ↑	Valid ↑	NFE
EQGAT-diff	–	$93.4_{\pm 0.21}$	$94.6_{\pm 0.24}$	500
SemlaFlow_{50}	–	$97.0_{\pm 0.21}$	$93.9_{\pm 0.12}$	50
Canon. SemlaFlow_{50}	1.0	$\mathbf{97.6}_{\pm 0.13}$	$94.6_{\pm 0.32}$	50
Canon. SemlaFlow_{50}	1.5	$97.4_{\pm 0.07}$	$\mathbf{94.7}_{\pm 0.12}$	50
Canon. SemlaFlow_{50}	2.0	$96.7_{\pm 0.04}$	$94.5_{\pm 0.03}$	50
SemlaFlow_{100}	–	$97.3_{\pm 0.08}$	$93.9_{\pm 0.19}$	100
Canon. SemlaFlow_{100}	1.0	$\mathbf{98.1}_{\pm 0.03}$	$95.0_{\pm 0.20}$	100
Canon. SemlaFlow_{100}	1.5	$97.8_{\pm 0.03}$	$95.0_{\pm 0.09}$	100
Canon. SemlaFlow_{100}	2.0	$97.5_{\pm 0.08}$	$\mathbf{95.1}_{\pm 0.07}$	100
Data	–	100.0	100.0	–

model is not conditioned on an additional symmetry-breaking variable, so training-time equivariant OT only changes the endpoint coupling while preserving the inference prior marginal; in this case, OT is compatible with the aligned slice prior and can be viewed as an aligned coupling on the canonical space. The situation changes once a non-equivariant conditioning signal such as canonical-rank PE is introduced. In this case, the model learns under a joint law $\widetilde{\pi}_1^{\mathrm{tr}}(C, \widetilde{Z}_1)$ over the condition and the initial noise, and training–sampling consistency requires this law to match the inference law $\widetilde{\pi}_1^{\mathrm{inf}}(C, \widetilde{Z}_1)$. A naive combination of PE, aligned prior, and OT does not explicitly preserve this conditional joint law: OT may improve endpoint matching, but it can simultaneously destroy the alignment between C and $\widetilde{Z}_1$ that the sampler relies on. We therefore include this non-aligned variant as a negative ablation, highlighted in red in Table 13. Its weaker few-step performance supports our theoretical claim that OT remains beneficial only when the induced coupling is compatible with the conditioning joint law. Thus, combining an aligned prior with OT under S_N canonicalization requires preserving the joint law of the condition and the initial noise.

Table 13: Canonical SemlaFlow ablations with aligned prior on GEOM-DRUG under S_N canonicalization. We compare equivariant and non-equivariant backbones with aligned priors and OT. The red rows report a deliberately non-aligned variant that combines PE with OT without preserving the conditional start-time joint law $\widetilde{\pi}_1^{\mathrm{tr}}(C, \widetilde{Z}_1) = \widetilde{\pi}_1^{\mathrm{inf}}(C, \widetilde{Z}_1)$; its weaker few-step performance is consistent with the training–sampling consistency analysis.

Group	Model	PE	OT	Mol Stab ↑	Valid ↑	NFE
	EQGAT-diff	-	-	$93.4_{\pm 0.21}$	$94.6_{\pm 0.24}$	500
	SemlaFlow_{20}	-	equivariant	$95.3_{\pm 0.14}$	$93.0_{\pm 0.10}$	20
	SemlaFlow_{50}	-	equivariant	$97.0_{\pm 0.21}$	$93.9_{\pm 0.12}$	50
	SemlaFlow_{100}	-	equivariant	$97.3_{\pm 0.08}$	$93.9_{\pm 0.19}$	100
S_N	Canon. SemlaFlow_{20}	-	equivariant	$95.1_{\pm 0.09}$	$93.5_{\pm 0.17}$	20
S_N	Canon. SemlaFlow_{50}	-	equivariant	$97.3_{\pm 0.08}$	$94.4_{\pm 0.02}$	50
S_N	Canon. SemlaFlow_{100}	-	equivariant	$97.5_{\pm 0.10}$	$94.2_{\pm 0.11}$	100
S_N	Canon. SemlaFlow_{20}	True	equivariant	$92.6_{\pm 0.18}$	$91.7_{\pm 0.20}$	20
S_N	Canon. SemlaFlow_{50}	True	equivariant	$96.3_{\pm 0.03}$	$93.9_{\pm 0.07}$	50
S_N	Canon. SemlaFlow_{100}	True	equivariant	$97.5_{\pm 0.10}$	$94.2_{\pm 0.20}$	100
	Data			100.0	100.0	–

Ablation on the canonical ordering methods. Table 14 compares an atomic-number ordering baseline with the Fiedler-based recipe on QM9. At the same NFE, the latter improves molecule stability by 1.16 percentage points and validity by 2.91 points, while reducing Opt-RMSD by 0.094 Å.

Atomic-number ordering does not distinguish same-element atoms without additional tie-breaking, whereas Fiedler ordering uses geometric information.

Table 14: QM9 generation recipes at 100 NFE. Stability and validity are percentages. The configurations differ in both ordering and additional generation components.

Recipe	Atom stability	Mol. stability	Validity	Opt-RMSD (Å)
Atomic-number ordering	$99.9_{\pm 0.0}$	$98.48_{\pm 0.06}$	$96.67_{\pm 0.30}$	$0.264_{\pm 0.003}$
Fiedler ordering	$99.9_{\pm 0.0}$	$99.64_{\pm 0.04}$	$99.58_{\pm 0.06}$	$0.17_{\pm 0.00}$

F SAMPLES FROM CANONICAL DIFFUSION

This section presents samples from Canonicalized SemlaFlow trained on GEOM-DRUG. The samples were generated randomly, but we have rotated them where necessary to aid visualization.

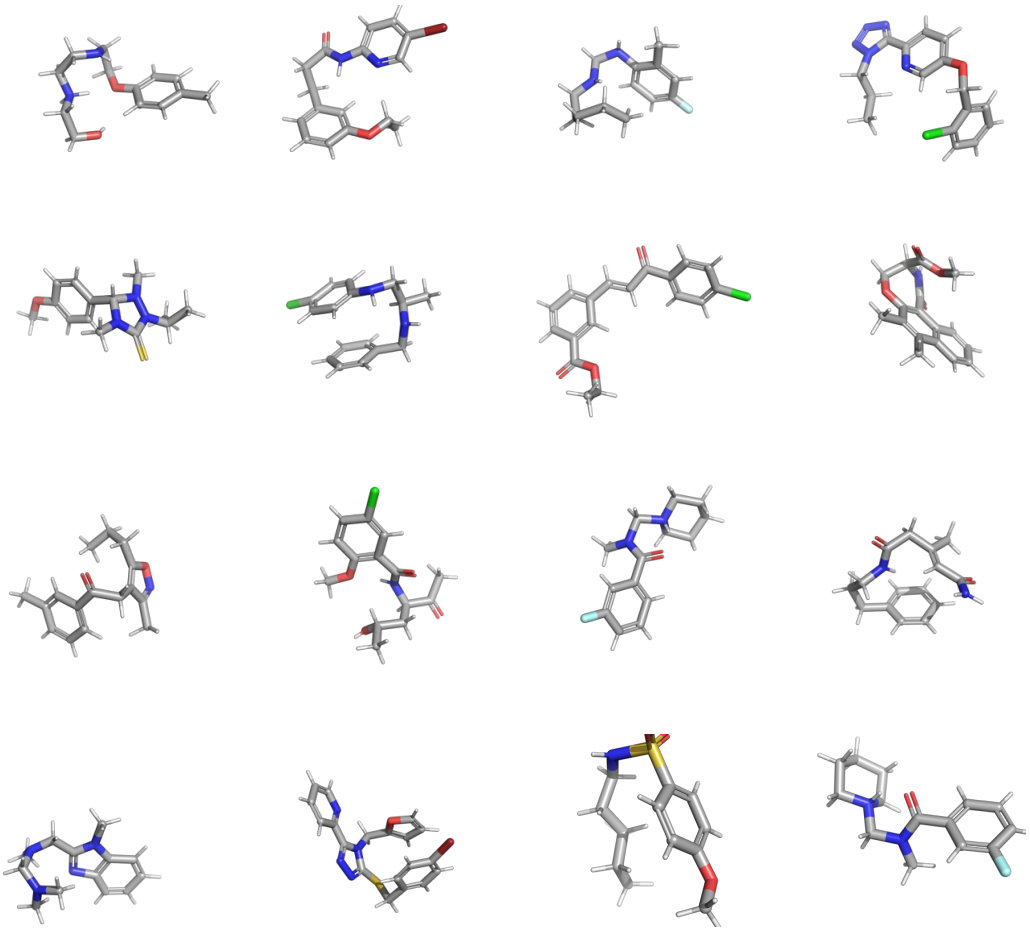


Figure 8: Random samples from Canonicalized SemlaFlow model trained on GEOM-DRUG. These samples were generated using 100 ODE integration steps.