FLDWM: WORLD MODEL FOR LEARNING APPROXIMATE SO(3) EQUIVARIANCE FOR SCALABLE AND GENERALIZABLE MOLECULE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariant neural networks provide physically accurate representations for 3D molecular generation but introduce computational bottlenecks that limit scalability, while scalable voxel-based methods lack explicit rotational symmetries. To bridge this gap, we introduce the Forced Latent Denoising World Model (FLDWM), the first world-model-based framework for 3D molecule generation. Operating on compressed voxel representations, FLDWM formulates symmetry learning as a latent dynamics problem: the model is trained to invert random SO(3) rotation augmentations, conditioned on the applied rotation via Feature-wise Linear Modulation (FiLM), which we show theoretically forces the denoiser to learn approximate SO(3) equivariance, bounded only by voxel discretization error. On GEOM-Drugs, FLDWM significantly outperforms state-of-the-art non-equivariant (VoxMol, NEBULA, RADM) and equivariant (EDM, MiDi) baselines on stability and validity while retaining fast inference, and demonstrates zero-shot generalization to the out-of-distribution PCQM dataset without fine-tuning. As a practical illustration, seed-conditioned generation from five FDA-approved drugs yields novel lead-optimization-range analogs (ECFP4 $T_c \in [0.4, 0.7]$), showing FLDWM's potential to propose alternative drug candidates.

1 INTRODUCTION

In-silico molecular generation is a fundamental challenge in molecular design, particularly in drug discovery. Following the success of AlphaFold Jumper et al. (2021), deep learning has emerged as a powerful paradigm for molecular modeling, generation, and design Simonovsky and Komodakis (2018); Gebauer et al. (2021); Gasteiger et al. (2020); Simm et al. (2021); Watson et al. (2023).

Molecules naturally exist in three-dimensional space and are best represented by atomic coordinates, atom types, bond types, and conformations, motivating direct generation in 3D. Point-cloud approaches apply diffusion models to atomic coordinates Hoogeboom et al. (2022); Xu et al. (2022); Igashov et al. (2022); Schneuing et al. (2022); Vignac et al. (2023), whereas voxel-based methods first discretize structures into volumetric grids before applying generative models Pinheiro et al. (2023); Nowara et al. (2024).

Molecular structures exhibit rotational and translational symmetries, with rotations described by the special orthogonal group $SO(3)$. Several works enforce these symmetries through explicitly equivariant networks Thomas et al. (2018); Hoogeboom et al. (2022); Vignac et al. (2023), which preserve physically meaningful transformation behavior and often improve sample efficiency, but introduce computational overhead that can limit scalability on modern hardware Nowara et al. (2025). In contrast, non-equivariant methods Pinheiro et al. (2023); Nowara et al. (2024) scale efficiently but rely on data augmentation to learn rotational structure implicitly, offering limited insight into how equivariance emerges.

World models learn latent state representations together with the dynamics governing transitions between states, and have achieved remarkable success in reinforcement learning and control Ha and Schmidhuber (2018); Hafner et al. (2019; 2023). Inspired by these advances, we formulate rotational symmetry learning as a latent dynamics problem and introduce the *Forced Latent Denoising World Model (FLDWM)*, which treats a molecular latent representation as the state and rotational

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.

augmentations as actions. The model is trained to predict the action required to recover the original latent state from its rotated counterpart, thereby learning approximate rotational equivariance within the latent space. Unlike explicitly equivariant architectures, this approach preserves the scalability of conventional voxel-based CNNs while encouraging equivariant representations through a world-model objective. Experiments demonstrate that FLDWM consistently improves molecular generation quality over both non-equivariant baselines — NEBULA Nowara et al. (2024), VoxMol Pinheiro et al. (2023) and RADM Ding and Hofmann (2025) — and equivariant ones, EDM Hoogeboom et al. (2022) and MiDi Vignac et al. (2023), and exhibits strong generalization to out-of-distribution molecular distributions.

2 Related Work

2.1 3D Molecule Generation

EDM Hoogeboom et al. (2022) introduced equivariant diffusion Ho et al. (2020) for molecular point clouds and was the first method to scale to GEOM-Drugs Axelrod and Gómez-Bombarelli (2022), while MiDi Vignac et al. (2023) extends this by jointly modeling 3D coordinates and molecular connectivity. A complementary line of work employs voxel-based representations Wang et al. (2022); Skalic et al. (2019); Ragoza et al. (2020); Pinheiro et al. (2023); Nowara et al. (2024), with VoxMol Pinheiro et al. (2023) and NEBULA Nowara et al. (2024) operating directly on voxelized atomic densities. Closest to our setting, RADM Ding and Hofmann (2025) likewise pursues scalable non-equivariant generation by handling rotations explicitly, through rotational alignment. FLDWM differs in learning to *invert* a known rotation as a world-model action, which is what yields the approximate equivariance certified in Sec. 3.4.

3 Method

Let $\mathbf{x} \sim p_{\mathrm{data}}$ be a molecule, represented as a continuous atomic density signal $x(u) \in \mathcal{X}(\Omega)$. We discretize this signal into a multi-channel voxel grid via a voxelization operator $\mathcal{V}_h$, where h is the spatial grid spacing Pinheiro et al. (2023). To learn rotational symmetry, we apply random spatial rotations R to the molecular coordinates prior to voxelization. These rotations are sampled uniformly from the 3D rotation group $SO(3)$ using Shoemake's algorithm Shoemake (1992), ensuring they are drawn from the Haar measure μ_{Haar}. Additional details are provided in the appendix.

3.1 Background

Following Nowara et al. (2024), we compress the voxelized molecules into a discrete latent representation using a VQ-VAE van den Oord et al. (2017) encoder $\mathcal{E}$, and subsequently apply a channel-wise normalization ψ_μ based on the empirical latent posterior statistics of the training set. We denote this entire composite encoding pipeline as $\Phi = \psi_\mu \circ \mathcal{E} \circ \mathcal{V}_h$. Thus, the canonical latent representation of a given molecule is defined as $z \coloneqq \Phi(\mathbf{x})$. Complete architectural and training details are provided in the appendix.

3.2 Forced Latent Denoising World Model (FLDWM)

The Forced Latent Denoising World Model (FLDWM), denoted as f_θ, operates as a World model designed to simultaneously remove noise from latent representations and invert random rotation augmentations.

Let T_R denote the left regular representation of the spatial rotation acting on the molecule, where $(T_R\, x)(u) = x(R^{-1}u)$. The induced transformation in the latent space is denoted by $\widetilde{T}_R$, such that the augmented (rotated) latent representation is given by $\widetilde{T}_R(\Phi(\mathbf{x})) \coloneqq \Phi(T_R\, \mathbf{x})$.

To formulate the denoising task, isotropic Gaussian noise is injected into the normalized rotated latent:

$$z_{\mathrm{noisy}} = \widetilde{T}_R(z) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{1}$$

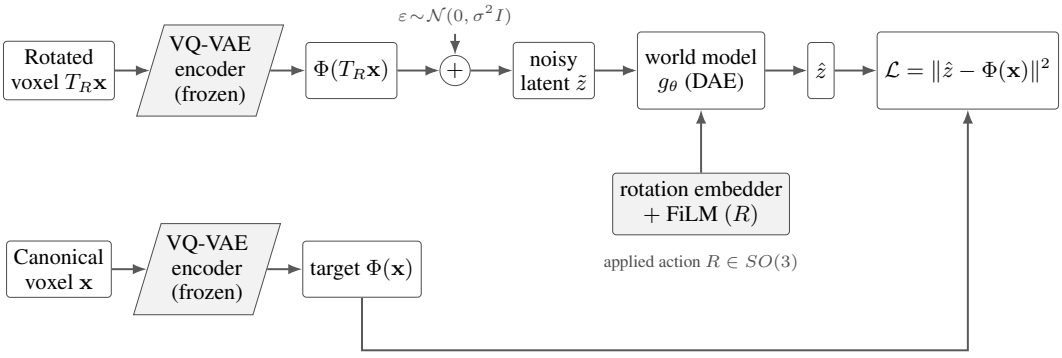

Figure 1: Overview of the Forced Latent Denoising World Model. A molecule is voxelized in a randomly rotated pose $T_R\mathbf{x}$ and in its canonical pose $\mathbf{x}$, and both are mapped to the latent space by the same frozen VQ-VAE encoder. Gaussian noise is added to the rotated latent, and the world model g_θ — conditioned on the applied action R through a rotation embedder and FiLM layers — is trained to recover the *canonical* latent. The model must therefore undo the rotation as well as the noise; it is this forced recovery of a known action that induces the approximate $SO(3)$ equivariance analysed in Sec. 3.4.

World-model formulation. We interpret f_θ as a single-step *world model* whose state space is the latent manifold $\mathcal{Z}$ and whose dynamics are governed by the rotation group action. The episode begins in the perturbed *initial state*

$$s_0 \;\coloneqq\; z_{\text{noisy}} \;=\; \widetilde{T}_R(z) + \epsilon, \tag{2}$$

that is, the canonical code $z = \Phi(\mathbf{x})$ after the rotation augmentation $\widetilde{T}_R$ and additive noise have displaced it from its canonical orientation. Observing the applied rotation through the conditioning signal $R_{\text{flat}} \in \mathbb{R}^9$, the model selects the *rotation-inversion action*

$$a_R \;\coloneqq\; \widetilde{T}_{R^{-1}}, \tag{3}$$

the unique latent transformation that undoes $\widetilde{T}_R$ (guaranteed to exist since $R^{-1} = R^\top \in SO(3)$). The learned transition dynamics f_θ then apply this action to advance the system to the terminal *canonical state*

$$s_1 \;=\; f_\theta\big(s_0,\, R_{\text{flat}}\big) \;\approx\; \Phi(\mathbf{x}) \;=\; z. \tag{4}$$

In this sense FLDWM is a world model of the $SO(3)$ symmetry: rather than merely denoising, it predicts how the latent state *evolves* when the inverse-rotation action is applied, recovering the canonical orientation regardless of the initial pose. As shown in Sec. 3.4, this action is structurally realized by the conditioning decomposition $f_\theta(\cdot,\, R_{\text{flat}}) = \widetilde{T}_{R^{-1}} \circ g_\theta(\cdot)$, where the FiLM layers implement the action a_R and g_θ performs the denoising.

The network f_θ receives this noisy latent alongside the flattened rotation matrix, $R_{\text{flat}} \in \mathbb{R}^9$, as the conditional information necessary to perform the rotational inversion. The model is optimized to predict the clean, canonical latent tensor $z = \Phi(\mathbf{x})$ by minimizing the following objective function:

$$\mathcal{L}(\theta) = \mathbb{E}_{R\sim\mu_{\text{Haar}}}\, \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\, \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)} \big\| f_\theta\big(\widetilde{T}_R(\Phi(\mathbf{x})) + \epsilon,\; R_{\text{flat}}\big) - \Phi(\mathbf{x}) \big\|^2 \tag{5}$$

3.3 World Model Conditioning

The model f_θ injects the rotation condition R_{flat} into the network's feature maps via Feature-wise Linear Modulation (FiLM). The flattened 3×3 rotation matrix is first embedded using a 5-layer MLP:

$$c = \text{MLP}(R_{\text{flat}}) \in \mathbb{R}^{3072}$$

At each residual block, this embedding is projected into scale (γ) and shift (β) vectors. The architecture utilizes an **unbounded** linear FiLM (omitting standard $\tanh$ activations for stronger gradient

flow) applied directly after group normalization. For an input feature map F, the exact parameter regression and feature modulation are formalized as:

$$[\gamma, \beta] = \text{Linear}(c) \qquad \text{(Parameter Regression)}$$
$$\hat{F} = \text{GroupNorm}(F) \odot (1 + \gamma) + \beta \qquad \text{(Feature Modulation)}$$

Result: By linearly scaling and shifting the mean and variance of the normalized feature maps based on the physical rotation, the network learns to decompose into $f_\theta(\cdot, R_{\text{flat}}) = \widetilde{T}_{R^{-1}} \circ g_\theta(\cdot)$, where g_θ acts as the denoiser backbone. This structural conditioning forces the network to internalize the $SO(3)$ symmetry group, applying the inverse latent rotation $\widetilde{T}_{R^{-1}}$ to "mentally rotate" the perturbed latent representation back to its canonical orientation $\Phi(\mathbf{x})$.

3.4 Proof of Approximate $SO(3)$ Equivariance via World Modelling

Equivariance. Let G be a group acting on a space through a representation ρ. A map f is G-*equivariant* if

$$f\big(\rho(g)\, x\big) = \rho(g)\, f(x), \qquad \forall\, g \in G. \tag{6}$$

When $G{=}\,SO(3)$, the group of 3D rotations, equivariance ensures that rotating a molecular input and then predicting yields the same result as predicting and then rotating—a fundamental physical symmetry.

$SO(3)$ **and the guaranteed existence of inverses.** The special orthogonal group $SO(3) = \{R \in \mathbb{R}^{3\times 3} \mid R^\top R = I,\, \det(R) = 1\}$ is a compact matrix Lie group, hence a subgroup of $GL(3, \mathbb{R})$. Every $R \in SO(3)$ satisfies $\det(R) = 1 \neq 0$, so R is invertible; the orthogonality constraint further gives the closed-form inverse $R^{-1} = R^\top$. This guarantees that the inverse of any rotation *always exists, belongs to* $SO(3)$*, and is trivially computable*. The induced signal-space map T_R (Sec. 3.4, FLDWM) is bijective with $(T_R)^{-1} = T_{R^{-1}}$, verified by $(T_{R^{-1}} \circ T_R)\, x(u) = x(R\, R^{-1}u) = x(u)$.

Haar-uniform rotation augmentation. To cover the full continuous group $SO(3)$—not merely a finite set of axis-aligned rotations—we sample $R \sim \mu_{\text{Haar}}$. The Haar measure μ_{Haar} is the *unique* bi-invariant probability measure on $SO(3)$; every rotation is equally likely. This ensures the training objective enforces equivariance for *all* rotations simultaneously.

Why the equivariance is approximate. The molecular density $x(u) = \sum_i k(\|u - r_i\|)$ with isotropic kernel k satisfies *exact* continuous equivariance: $x_R(u) = x(R^{-1}u)$, since $\|u - Rr_i\| = \|R^{-1}u - r_i\|$ by orthogonality of R. However, in practice the network operates on a discrete voxel grid Λ_h with spacing h. For generic $R \in SO(3)$, the pre-image $R^{-1}u_{ijk}$ falls *between* grid points—the cubic lattice is preserved only by the finite octahedral rotation group O (24 proper rotations), which has Haar measure zero in $SO(3)$. Consequently, the sampling and rotation operators do not commute: $\mathcal{S}_h \circ T_R \neq T_R^\Lambda \circ \mathcal{S}_h$. The voxelizer itself introduces no error: it evaluates the kernel at the exact continuous atom–voxel distance, so $\mathcal{V}_h(T_R\mathbf{x})$ is an *exact* sampling of the rotated density, as is $\mathcal{V}_h(\mathbf{x})$ of the unrotated one. What the grid costs is that these are samplings of the same field on two *incommensurable* lattices, Λ_h and $R^{-1}\Lambda_h$: for $R \notin O$ no operation on the grid carries one to the other, so the rotated configuration is only partly recoverable from the unrotated samples. The size of that loss is the error of the best available reconstruction. Reconstructing trilinearly, the value at $R^{-1}u_{ijk}$ is a convex combination of the eight corners $v_1, \ldots, v_8$ of the enclosing cell, each within one cell diagonal, so for L-Lipschitz x

$$\big|x(R^{-1}u_{ijk}) - x_{\text{tri}}(u_{ijk})\big| \;\le\; \max_j L\, \|R^{-1}u_{ijk} - v_j\| \;\le\; L\, h\sqrt{3} \;=\; \mathcal{O}(h). \tag{7}$$

For a C^2 density this sharpens to $\frac{3}{8}h^2 M_2$ with $M_2 = \max_i \sup|\partial_i^2 x|$; empirically the relative reconstruction error scales as $h^{1.55}$ on GEOM-Drugs (appendix), between the two rates, as expected for a density whose atomic width (0.5 Å) is only twice the grid spacing. This error is *irreducible* at fixed grid resolution, is independent of model architecture, and vanishes as $h \to 0$.

Training objective. Recall that the world model f_θ, conditioned on the flattened rotation matrix $R_{\text{flat}} \in \mathbb{R}^9$ via FiLM layers, is trained to minimize the objective in Eq. equation 5.

Theorem 1 (Approximate $SO(3)$ equivariance)**.** *Suppose f_{θ^*} achieves $\mathcal{L}(\theta^*) \le \delta^2$. Decompose the conditioned model as $f_{\theta^*}(\,\cdot\,, R_{\text{flat}}) = \widetilde{T}_{R^{-1}} \circ g_{\theta^*}(\,\cdot\,)$, where g_{θ^*} is the denoiser backbone and $\widetilde{T}_{R^{-1}}$ is the inverse latent rotation learned via FiLM conditioning on R. Then g_{θ^*} satisfies $SO(3)$-equivariance on the image of Φ:*

$$g_{\theta^*}\big(\widetilde{T}_R(z)\big) \;=\; \widetilde{T}_R\big(g_{\theta^*}(z)\big), \qquad \textit{with error } 2\delta\,(1+\varepsilon_h),\;\; \varepsilon_h = \mathcal{O}(h), \tag{8}$$

where $z \coloneqq \Phi(\mathbf{x})$.

Proof. (i) The world model learns $\widetilde{T}_{R^{-1}}$. Write $z = \Phi(\mathbf{x})$. The loss equation 5 requires the full model, given the rotated latent $\widetilde{T}_R(z)$ and the conditioning R_{flat}, to recover the canonical code z. Since $\widetilde{T}_R$ is a bijection on $\mathcal{Z}$ (with inverse $\widetilde{T}_{R^{-1}}$, guaranteed by $R^{-1} = R^\top \in SO(3)$), the *unique* transformation that maps $\widetilde{T}_R(z)$ back to z for every z is $\widetilde{T}_{R^{-1}}$. Without learning this transformation, the loss cannot decrease. The convergence $\mathcal{L}(\theta^*) \le \delta^2$ therefore certifies that the FiLM layers, conditioned on R, have learned to apply $\widetilde{T}_{R^{-1}}$ to the backbone output $g_{\theta^*}(\widetilde{T}_R(z))$.

(ii) From convergence. Step (i) fixes the *form* of the trained model; the objective now tells us how well it performs. The two combine as follows. Evaluating the model on the state $s_0 = z_{\text{noisy}}$ it is trained on and substituting the decomposition,

$$f_{\theta^*}\big(z_{\text{noisy}},\, R_{\text{flat}}\big) \;=\; \widetilde{T}_{R^{-1}}\Big(g_{\theta^*}\big(z_{\text{noisy}}\big)\Big), \qquad z_{\text{noisy}} = \widetilde{T}_R(z) + \epsilon,$$

so the FiLM factor moves outside the backbone and the objective equation 5 can be rewritten entirely in terms of g_{θ^*}:

$$\mathcal{L}(\theta^*) \;=\; \mathbb{E}_{R\sim\mu_{\text{Haar}}}\,\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\,\mathbb{E}_{\epsilon}\Big\|\,\widetilde{T}_{R^{-1}}\Big(g_{\theta^*}\big(z_{\text{noisy}}\big)\Big) - z\,\Big\|^2.$$

The hypothesis $\mathcal{L}(\theta^*) \le \delta^2$ is therefore precisely the statement that the world model, after applying its rotation-inversion action, returns the canonical code up to a residual of size $\delta \coloneqq \sqrt{\mathcal{L}(\theta^*)}$ in root mean square over $(R, \mathbf{x}, \epsilon)$. Writing that residual per sample, and suppressing ϵ in the notation, gives the optimality condition used throughout the rest of the proof:

$$\big\|\,\widetilde{T}_{R^{-1}}\big(g_{\theta^*}(\widetilde{T}_R(z))\big) - z\,\big\| \;\le\; \delta. \tag{9}$$

Eq. equation 9 is the bridge between training and geometry: everything that follows is obtained from it by applying $\widetilde{T}_R$ and comparing two instances of the same bound.

(iii) Identity case. Setting $R = I$ in Eq. equation 9 and using $\widetilde{T}_I = \mathrm{Id}$ gives

$$\big\|g_{\theta^*}(z) - z\big\| \;\le\; \delta. \tag{10}$$

Note this is a bound, not an identity: the objective equation 5 injects noise of scale σ, and the minimiser of a square loss under noise is the conditional mean $\mathbb{E}[\,z \mid z+\epsilon\,]$, which is not the identity map. Eq. equation 10 is what convergence delivers, and it is what Eq. equation 16 uses.

(iv) Equivariance. Apply $\widetilde{T}_R$ to the vector appearing inside Eq. equation 9. Because $\widetilde{T}_R$ is linear and $\widetilde{T}_R\widetilde{T}_{R^{-1}} = \mathrm{Id}$ (group inverse property), the first term collapses and the second picks up $\widetilde{T}_R$:

$$\widetilde{T}_R\Big[\,\widetilde{T}_{R^{-1}}\big(g_{\theta^*}(\widetilde{T}_R z)\big) - z\,\Big] \;=\; g_{\theta^*}\big(\widetilde{T}_R(z)\big) - \widetilde{T}_R(z).$$

Taking norms and using $\|\widetilde{T}_R v\| \le (1+\varepsilon_h)\|v\|$ from Eq. equation 14 together with Eq. equation 9,

$$\big\|g_{\theta^*}\big(\widetilde{T}_R(z)\big) - \widetilde{T}_R(z)\big\| \;\le\; (1+\varepsilon_h)\,\delta. \tag{11}$$

So the backbone fixes the rotated code to within $(1+\varepsilon_h)\delta$. This is the statement at R; Eq. equation 10 is the same statement at $R = I$. Inserting $\widetilde{T}_R(z)$ and applying the triangle inequality to the two of them,

$$\big\|g_{\theta^*}\big(\widetilde{T}_R(z)\big) - \widetilde{T}_R\big(g_{\theta^*}(z)\big)\big\| \;\le\; \underbrace{\big\|g_{\theta^*}(\widetilde{T}_R z) - \widetilde{T}_R z\big\|}_{\le (1+\varepsilon_h)\delta \text{ by } equation\ 11} \;+\; \underbrace{\big\|\widetilde{T}_R\big(z - g_{\theta^*}(z)\big)\big\|}_{\le (1+\varepsilon_h)\delta \text{ by } equation\ 10}, \tag{12}$$

which is the $SO(3)$-equivariance condition equation 6 up to a residual: the denoiser backbone commutes with the latent rotation to within the bound of Eq. equation 16. Two residuals are spent, one at R and one at the identity, which is the origin of the factor two below.

(v) Error budget. The equivariance equation 12 holds only approximately; we now bound the residual.

(a) Training residual. By hypothesis $\mathcal{L}(\theta^*) \le \delta^2$, so for each $(R, \mathbf{x})$:

$$\left\| f_{\theta^*}\big(\widetilde{T}_R(z),\, R_{\text{flat}}\big) - z \right\| \;\le\; \delta. \tag{13}$$

Equivalently, $\|\widetilde{T}_{R^{-1}}(g_{\theta^*}(\widetilde{T}_R(z))) - z\| \le \delta$. Since $\widetilde{T}_R$ is an isometry (rotation preserves norms), left-multiplying by $\widetilde{T}_R$ preserves the bound: $\|g_{\theta^*}(\widetilde{T}_R(z)) - \widetilde{T}_R(z)\| \le \delta$.

(b) Discretization error. The latent action of Sec. 3.2 is the one the pipeline implements, $\widetilde{T}_R(\Phi(\mathbf{x})) \coloneqq \Phi(T_R\mathbf{x})$, so no gap arises between encoding a rotated molecule and rotating an encoded one — that difference is zero by construction. Discretization enters instead through the isometry invoked in (a), which the grid satisfies only approximately. Write ε_h for the norm defect,

$$\varepsilon_h \;\coloneqq\; \sup_{\mathbf{x},R} \left| \tfrac{\|\widetilde{T}_R(\Phi(\mathbf{x}))\|}{\|\Phi(\mathbf{x})\|} - 1 \right|, \qquad \text{so that} \qquad \|\widetilde{T}_R v\| \le (1+\varepsilon_h)\|v\|. \tag{14}$$

At the voxel level this defect is a *quadrature* error, not a reconstruction error: $\|\mathcal{V}_h(T_R\mathbf{x})\|^2 = \sum_{u\in\Lambda_h} |x(R^{-1}u)|^2$ and $\|\mathcal{V}_h(\mathbf{x})\|^2 = \sum_{u\in\Lambda_h} |x(u)|^2$ are midpoint sums of the same integrand $|x|^2$ over the two lattices $R^{-1}\Lambda_h$ and Λ_h, both of spacing h. Each approximates $h^{-3}\int |x|^2$ with an error bounded by the variation of $|x|^2$ across one cell, so their difference obeys

$$\big|\, \|\mathcal{V}_h(T_R\mathbf{x})\|^2 - \|\mathcal{V}_h(\mathbf{x})\|^2 \,\big| \;\le\; 2\,h\sqrt{3}\,\|\nabla|x|^2\|_{L^1} \;=\; \mathcal{O}(h), \tag{15}$$

and propagating through the $L_{\mathcal{E}}$-Lipschitz encoder leaves $\varepsilon_h = \mathcal{O}(h)$. We stress that Eq. equation 15 bounds the voxel-level contribution only; ε_h also absorbs any non-equivariance of the learned encoder, which is not a function of h and which we therefore report as a measured quantity rather than bound analytically.

Combined bound. Adding the two bracketed terms of Eq. equation 12,

$$\left\| g_{\theta^*}\big(\widetilde{T}_R(z)\big) - \widetilde{T}_R\big(g_{\theta^*}(z)\big) \right\| \;\le\; 2\,\delta\,(1+\varepsilon_h) \;=\; 2\delta \;+\; \mathcal{O}(h). \tag{16}$$

The first term is reducible with model capacity and optimisation; the second is irreducible at fixed grid resolution and vanishes as $h \to 0$. The factor two is structural, not an artefact of loose bounding: one residual is spent evaluating g_{θ^*} at the rotated code and one on rotating its output at the identity. Empirically (appendix) $\delta = 0.1957$ and $\varepsilon_h = 0.018$ measured directly on 10,000 held-out GEOM-Drugs molecules, giving a certificate of 0.398 against a measured equivariance residual of 0.1607 ± 0.0001 — satisfied with a margin of $2.5\times$. Restricting the rotations to the lattice-preserving subgroup O, where the discretization term vanishes identically, changes the measured residual by about 1% (0.1589): the grid accounts for only about one percent of the certificate. The remainder tracks δ, which reflects how far optimisation has converged rather than a limit of the construction; the discretization term is the only contribution intrinsic to the representation, and it vanishes as $h \to 0$. □

3.5 Empirical Quantification of the Error Budget

Theorem 1 bounds the equivariance error by $2\delta(1+\varepsilon_h)$, and both constants are measurable at the operating point of our experiments. We evaluate the trained checkpoints on 10,000 held-out GEOM-Drugs molecules $\times$ 8 Haar-random rotations (80,000 pairs) at $h = 0.25$ Å (64^3 voxel grid, 8^3 latent grid, $\sigma = 1.8$); uncertainties are standard errors of the mean. The protocol is given in Appendix A.4.

All entries are relative ℓ_2 errors, so the certificate and the residual are directly comparable; Appendix A.4 states the conversion from the per-element convention in which $\mathcal{L}$ is reported.

Three observations. First, the bound holds with room to spare: the measured residual is 40% of the certificate, a margin of $2.5\times$. Second, the budget is dominated by δ, not by the grid — restricting the rotations to the octahedral subgroup O, where the lattice action is exact and the discretization

Table 1: The error budget of Theorem 1 at the resolution used in the experiments. The certificate is satisfied with a margin of $2.5\times$, and the residual is an order of magnitude below the no-model baseline.

Quantity	Meaning	Measured
$\|z\|$	scale of the normalised latent, per element	0.7916 ± 0.0028
δ	$\sqrt{\mathcal{L}(\theta^*)}$, training residual	0.1957
ε_h	isometry defect, Eq. equation 14	0.018 ± 0.0004
certificate	$2\delta(1+\varepsilon_h)$, Eq. equation 16	0.398
equivariance residual	LHS of Eq. equation 8, Haar	$\mathbf{0.1607 \pm 0.0001}$
restricted to O	lattice-preserving rotations	0.1589 ± 0.0001
baseline	$\|\Phi(T_R\mathbf{x}) - \Phi(\mathbf{x})\|$, no model	1.9516 ± 0.0016

contribution vanishes identically, changes the residual by about 1% (0.1589 vs 0.1607). Third, the effect is large relative to doing nothing: without the world model the two codes differ by 1.95, so the rotation-inversion action removes roughly 92% of the pose dependence — the residual is $12\times$ smaller than the quantity it is suppressing. Separately, sweeping the grid spacing confirms the rate assumed in Eq. equation 7: the voxel-level reconstruction error falls as $h^{1.55}$ over $h \in [0.125, 1.0]$ Å (Table 4), i.e. at least as fast as the $\mathcal{O}(h)$ the analysis requires.

4 Experiments

4.1 Dataset

FLDWM is trained on GEOM-DRUGS Axelrod and Gómez-Bombarelli (2022) and its cross-dataset generalization is evaluated on PCQM Nakata and Maeda (2023); Nakata and Shimazaki (2017); Nakata et al. (2020). We adopt the MiDi Vignac et al. (2023) train/validation/test split (1.1M/146K/146K molecules), sub-sampling 100k molecules per training epoch.

4.2 Baselines and Metrics

Following NEBULA (Nowara et al., 2024), we perform seed-conditioned generation. We compare against non-equivariant baselines — NEBULA, VoxMol (Pinheiro et al., 2023) and RADM (Ding and Hofmann, 2025) — and equivariant baselines EDM (Hoogeboom et al., 2022) and MiDi (Vignac et al., 2023). We report the metrics of Vignac et al. (2023) and Pinheiro et al. (2023), and additionally the Tanimoto similarity between each seed and its generated molecule (Nowara et al., 2024) (metric details in Appendix A.6).

4.3 Molecule Generation

As in NEBULA (Nowara et al., 2024), we generate 10K molecules from 1K seeds by taking a variable number of WJS steps in the noisy latent space followed by a denoising jump, and sanitize outputs with RDKit (Landrum, 2016) (see appendix). FLDWM surpasses both VoxMol and NEBULA on several metrics while remaining competitive on the rest (Tables 1 and 2; GEOM-Drugs and PCQM). It is also scalable, averaging only 0.66 seconds per molecule (0.21 seconds for 10-step WJS) despite its 213.31M parameters, compared to 12.56M for the NEBULA DAE.

None of the three diffusion baselines (RADM, MiDi, EDM) is natively seed-conditioned. We therefore condition them by partial noising, following Meng et al. (2022): the seed is corrupted by the model's own forward process to $t_{\text{start}} = \text{round}(\text{NF} \cdot T)$, and the unmodified reverse process is run back to 0. The noise fraction $\text{NF} \in (0, 1]$ is the sole perturbation parameter, fixing the denoiser's input signal-to-noise ratio $\bar{\alpha}_t^2/\bar{\sigma}_t^2$ at t_{start}.

The WJS step count k and the noise fraction NF are not the same control parameter, so we do not equate them directly. We instead match them on the signal-to-noise ratio of the state entering the denoiser, measured for FLDWM and inverted through each baseline's own schedule: this gives $\text{NF} = 0.77$ (RADM, EDM) and 0.885 (MiDi) on GEOM-Drugs, and 0.80 and 0.90 on PCQM.

Table 2: Seeded generation results on **GEOM**. MiDi, RADM, and EDM are reported only at their SNR-matched noise fractions; FLDWM, VoxMol, and NEBULA are evaluated across WJS step counts. All results are averaged over 5 repeats.

WJS Steps	tan. sim. % ↑	stable sanit. % ↑	stable atom % ↑	valid % ↑	valency W_1 ↓	atom TV ↓	bond TV ↓	bond len W_1 ↓	bond ang W_1 ↓	avg. t [s/mol.] ↓
0 (*data*)	100.00	99.90	99.90	99.80	0.00	0.00	0.03	0.00	0.05	–
MiDi (NF 0.885)	7.28 (±0.03)	58.46 (±0.42)	95.99 (±0.03)	89.70 (±0.32)	0.09 (±0.00)	0.07 (±0.00)	0.15 (±0.00)	0.00 (±0.00)	4.46 (±0.01)	–
RADM (NF 0.77)	10.00 (±0.02)	61.86 (±0.20)	95.69 (±0.03)	**99.53** (±0.06)	0.15 (±0.00)	0.04 (±0.00)	0.01 (±0.00)	0.00 (±0.00)	0.76 (±0.01)	–
EDM (NF 0.77)	6.99 (±0.02)	42.20 (±0.34)	91.59 (±0.08)	98.40 (±0.07)	0.24 (±0.00)	0.10 (±0.00)	0.01 (±0.00)	0.00 (±0.00)	5.35 (±0.02)	–
5 *VoxMol*	80.84 (±0.96)	85.54 (±0.15)	99.43 (±0.04)	90.12 (±0.36)	0.26 (±0.00)	0.02 (±0.00)	0.03 (±0.00)	0.00 (±0.00)	0.67 (±0.02)	0.38
10	71.44 (±1.82)	85.71 (±0.36)	99.36 (±0.04)	89.86 (±0.28)	0.25 (±0.00)	0.02 (±0.00)	0.03 (±0.00)	0.00 (±0.00)	0.67 (±0.03)	0.66
50	44.99 (±1.03)	86.42 (±0.94)	99.35 (±0.02)	90.15 (±0.76)	0.25 (±0.00)	0.03 (±0.00)	0.03 (±0.00)	0.00 (±0.00)	0.54 (±0.02)	–
100	35.18 (±0.34)	**87.44** (±1.06)	99.37 (±0.02)	90.52 (±0.81)	0.25 (±0.00)	0.03 (±0.00)	0.03 (±0.00)	0.00 (±0.00)	0.50 (±0.00)	–
10 *NEBULA*	**88.47** (±0.18)	85.73 (±0.12)	**99.46** (±0.01)	90.29 (±0.11)	0.26 (±0.00)	0.03 (±0.00)	0.03 (±0.00)	0.00 (±0.00)	0.77 (±0.00)	**0.21**
20	84.30 (±0.46)	85.46 (±0.31)	99.42 (±0.01)	89.82 (±0.31)	0.26 (±0.00)	0.03 (±0.00)	0.03 (±0.00)	0.00 (±0.00)	0.85 (±0.01)	0.23
50	63.85 (±0.46)	83.11 (±0.31)	99.16 (±0.01)	86.61 (±0.31)	0.25 (±0.00)	0.03 (±0.00)	0.03 (±0.00)	0.00 (±0.00)	1.15 (±0.02)	–
100	38.33 (±0.80)	79.52 (±0.85)	98.45 (±0.04)	81.42 (±0.75)	0.23 (±0.00)	0.04 (±0.00)	0.03 (±0.00)	0.00 (±0.00)	1.80 (±0.01)	–
10 **FLDWM**	88.39 (±0.00)	87.29 (±0.09)	98.20 (±0.03)	96.62 (±0.07)	**0.05** (±0.00)	**0.00** (±0.00)	**0.00** (±0.00)	0.00 (±0.00)	**0.37** (±0.00)	0.66
20	86.04 (±0.10)	84.85 (±0.11)	97.86 (±0.01)	96.43 (±0.05)	0.06 (±0.00)	0.02 (±0.00)	0.00 (±0.00)	0.00 (±0.00)	0.56 (±0.00)	0.91
50	56.18 (±0.26)	62.63 (±1.47)	91.92 (±2.33)	97.71 (±1.68)	0.19 (±0.04)	0.02 (±0.01)	0.01 (±0.00)	0.00 (±0.00)	2.18 (±1.09)	–
100	28.73 (±0.30)	47.16 (±0.26)	86.06 (±0.17)	98.01 (±0.08)	0.31 (±0.00)	0.06 (±0.00)	0.01 (±0.00)	0.01 (±0.00)	3.44 (±0.01)	–

Table 3: **Cross-dataset generalizability**: seeded generation results on **PCQM**. MiDi, RADM, and EDM are reported only at their SNR-matched noise fractions. FLDWM outperforms the baselines on most metrics. All results are averaged over 5 repeats of each experiment.

Method/setting	tan. sim. % ↑	stable sanit. % ↑	stable atom % ↑	valid % ↑	valency W_1 ↓	atom TV ↓	bond TV ↓	bond len W_1 ↓	bond ang W_1 ↓
0 (*data*)	100.00	95.51	99.62	99.40	2.98	0.90	0.24	0.01	3.40
MiDi (NF 0.90)	7.88 (±0.03)	78.25 (±0.21)	96.94 (±0.05)	97.12 (±0.09)	0.06 (±0.00)	0.13 (±0.00)	0.00 (±0.00)	0.01 (±0.00)	2.34 (±0.04)
RADM (NF 0.80)	7.87 (±0.01)	73.02 (±0.32)	96.73 (±0.04)	99.74 (±0.05)	0.10 (±0.00)	0.16 (±0.00)	0.00 (±0.00)	0.01 (±0.00)	2.68 (±0.01)
EDM (NF 0.80)	6.30 (±0.04)	49.62 (±0.42)	92.59 (±0.07)	99.10 (±0.10)	0.20 (±0.00)	0.15 (±0.00)	0.01 (±0.00)	0.00 (±0.00)	5.68 (±0.03)
5 *VoxMol*	10.95 (±1.85)	92.45 (±0.29)	95.60 (±0.07)	97.11 (±0.17)	0.21 (±0.00)	0.43 (±0.01)	0.14 (±0.00)	0.00 (±0.00)	4.04 (±0.07)
10	9.74 (±0.62)	92.08 (±0.83)	96.89 (±0.11)	96.49 (±0.63)	0.22 (±0.00)	0.35 (±0.00)	0.14 (±0.00)	0.00 (±0.00)	3.49 (±0.05)
50	9.75 (±0.38)	90.73 (±0.54)	98.44 (±0.07)	95.17 (±0.30)	0.24 (±0.00)	0.15 (±0.01)	0.13 (±0.00)	0.00 (±0.00)	2.36 (±0.04)
100	9.81 (±0.21)	89.93 (±0.47)	98.62 (±0.04)	94.81 (±0.43)	0.25 (±0.00)	0.10 (±0.01)	0.12 (±0.00)	0.00 (±0.00)	1.98 (±0.07)
200	9.86 (±0.22)	90.44 (±0.85)	98.77 (±0.07)	94.68 (±0.56)	0.25 (±0.00)	0.06 (±0.00)	0.12 (±0.00)	0.00 (±0.00)	1.62 (±0.05)
10 *NEBULA*	35.10 (±0.57)	95.18 (±0.26)	75.64 (±0.13)	98.86 (±0.23)	0.23 (±0.00)	0.84 (±0.00)	0.24 (±0.00)	(±0.01) (±0.00)	4.59 (±0.03)
20	32.19 (±0.38)	95.44 (±0.27)	76.43 (±0.18)	98.80 (±0.13)	0.22 (±0.00)	0.83 (±0.00)	0.23 (±0.00)	0.01 (±0.00)	4.72 (±0.05)
50	25.62 (±0.26)	95.64 (±0.26)	78.35 (±0.18)	98.36 (±0.12)	0.20 (±0.00)	0.79 (±0.00)	0.23 (±0.00)	0.01 (±0.00)	5.19 (±0.06)
100	18.38 (±0.27)	**96.16** (±0.19)	80.04 (±0.20)	97.88 (±0.35)	0.18 (±0.00)	0.74 (±0.00)	0.23 (±0.00)	0.01 (±0.00)	6.39 (±0.04)
200	12.58 (±0.09)	95.72 (±0.49)	79.77 (±0.51)	96.72 (±0.26)	0.19 (±0.01)	0.68 (±0.00)	0.23 (±0.00)	0.00 (±0.00)	9.30 (±0.09)
10 **FLDWM** (Ours)	**88.62** (±0.07)	95.94 (±0.08)	**99.63** (±0.01)	**99.95** (±0.03)	**0.01** (±0.00)	**0.01** (±0.00)	**0.00** (±0.00)	**0.00** (±0.00)	**0.50** (±0.00)
20	86.69 (±0.12)	93.96 (±0.13)	99.44 (±0.02)	99.84 (±0.02)	**0.01** (±0.00)	**0.01** (±0.00)	**0.00** (±0.00)	0.00 (±0.00)	0.61 (±0.00)
50	70.90 (±0.31)	79.26 (±0.25)	97.96 (±0.05)	98.56 (±0.18)	0.06 (±0.00)	0.02 (±0.00)	0.00 (±0.00)	0.00 (±0.00)	1.00 (±0.00)
100	41.04 (±0.16)	57.52 (±0.30)	94.96 (±0.07)	96.58 (±0.07)	0.14 (±0.00)	0.04 (±0.00)	0.01 (±0.00)	0.00 (±0.00)	1.73 (±0.01)
200	18.41 (±0.08)	43.41 (±0.11)	90.47 (±0.04)	94.24 (±0.21)	0.23 (±0.00)	0.08 (±0.00)	0.01 (±0.00)	0.00 (±0.00)	3.31 (±0.02)

4.4 Cross-Dataset Generalization

FLDWM significantly outperforms the baselines on cross-dataset transfer: by learning $SO(3)$ rotational symmetry, it generalizes to out-of-distribution data, achieving high stability and validity while closely matching the structural distributions of the test set. We attribute this to $SO(3)$ equivariance capturing universal rigid-body properties rather than features restricted to a single molecular distribution. As a practical demonstration of this transferable chemistry prior, seed-conditioned generation from five FDA-approved drugs (Bedaquiline, Osimertinib, Sotorasib, Imatinib, and Ivacaftor) yielded structurally novel, IP-defensible analogs in the lead-optimization regime (ECFP4 Tanimoto $T_c \in [0.4, 0.7]$; see Section A.7).

4.5 Ablation: Is the Explicit Rotation Hint Necessary?

Rotation augmentation alone might be thought to recover most of the benefit of the world-model formulation. It does not. Withholding R at matched capacity — ZERO-COND retains every FiLM layer but feeds it a constant zero vector (213.31M, identical to the conditioned model), while NO-FILM re-

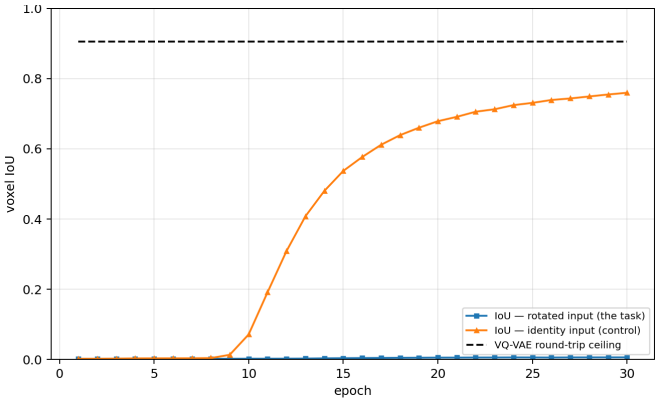

Figure 2: Withholding R at matched capacity (ZERO-COND): reconstruction IoU against the frozen VQ-VAE's 0.906 round-trip ceiling. The identity-input control climbs to 0.760 — the same weights denoise perfectly well on un-rotated inputs — while rotated-input IoU never leaves the axis ($0.0018 \rightarrow 0.0058$ over 30 epochs). The gap between the two curves is the evidence that the failure is informational rather than an optimisation artefact. Full diagnostics in Appendix A.5.

moves them outright (102.35M) — collapses both variants to the same place (Fig. 2): rotated-input loss stalls at 0.494 against a predict-the-mean baseline of 0.575, and reconstruction IoU never leaves the floor (0.0058, against the frozen autoencoder's 0.906 ceiling). Decisively, the *same weights* denoise un-rotated inputs normally (IoU 0.760), so the failure is specific to the rotation term rather than to under-training. This is the square-loss optimum once R is unobservable, since $\mathbb{E}[z_{\text{can}} \mid z_{\text{rot}}]$ averages over the orbit: withholding the hint renders the regression target ill-posed rather than merely unhinted. It is therefore the *forced* recovery of a known action that induces the approximate equivariance certified in Sec. 3.4. Appendix A.5 reports the full diagnostics and a further ablation of the 15% identity fraction, whose removal degrades the model further still (identity IoU $0.775 \rightarrow 0.0023$).

5 Conclusion

We introduced FLDWM, a world-model formulation for 3D molecule generation that treats a compressed voxel latent as the state and rotational augmentation as the action, and trains the model to recover the action required to return a rotated latent to its canonical pose. This *forced* recovery induces approximate $SO(3)$ equivariance in a conventional, scalable CNN backbone, without the architectural constraints of explicitly equivariant networks. We proved a bound on the residual equivariance defect and measured every constant entering it on held-out GEOM-Drugs molecules, finding the certificate satisfied with a factor of 2.5 to spare and the residual an order of magnitude below the no-model baseline. Empirically, FLDWM improves on non-equivariant voxel baselines for seeded generation and transfers to an out-of-distribution dataset; seed-conditioned generation from five FDA-approved drugs yields novel analogs in the lead-optimization regime (ECFP4 $T_c \in [0.4, 0.7]$). An ablation at matched capacity further shows that the explicit rotation hint is not an optimisation aid but the ingredient that makes the training objective well-posed.

Two directions remain open. Our aim was not to construct an exactly equivariant architecture, but to show that *forcing* the recovery of a known rotation action yields a world model that *learns* approximate $SO(3)$ equivariance while retaining the scalability and cross-dataset generalization that make voxel methods attractive. The equivariance we certify is approximate by construction: in the idealised setting the residual reduces to the voxel discretization term, while in practice the frozen encoder is not steerable and contributes a further component (Appendix A.4). An encoder built from steerable kernels would remove that component and tighten the certificate, making the approach more practical still — at the cost of reintroducing part of the architectural overhead that FLDWM is designed to avoid. Second, generation quality declines as the number of walk–jump steps grows. We attribute this to the sample drifting away from its seed rather than to a defect of the sampler — the same sampler improves with step count in voxel space — but characterising that trade-off, and the behaviour of the certificate along the sampler trajectory, remains open.

Ethics Statement

This work develops a generative model for small molecules trained on public conformer datasets (GEOM-Drugs, PCQM). No human subjects, personally identifiable information, or sensitive data are involved. Generated structures are computational hypotheses only: they carry no safety, synthesizability, or efficacy guarantee, and the lead-optimization analogs are presented as an illustration of the method, not as drug candidates. As with any molecular generator, the model could in principle be directed toward harmful compounds; we release no tooling specific to that use, and the seed-conditioned protocol requires an explicit user-supplied target.

Reproducibility Statement

All architectural, training, and evaluation details are given in the appendix: voxelization and encoder configuration, world-model architecture and FiLM conditioning, optimizer and schedule, and the generation procedure. The assumptions and constants entering Theorem 1 are stated in Sec. 3.4, and the appendix reports the protocol used to measure the equivariance residual, δ, and the discretization term ε_h. Source code and trained checkpoints will be released upon acceptance.

Use of Large Language Models

We used generative AI tools for coding and software engineering, for writing and polishing, and for checking proofs and the rigour of the analysis. The research questions, the model formulation, and all scientific conclusions are the authors'. All AI-assisted output was reviewed by the authors, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

A comparative study of SMILES, SELFIES, and ECFP4 representations for molecular similarity search. ChemRxiv, 2024.

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

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 2018. URL https://proceedings.mlr.press/v75/cheng18a.html.

Yuhui Ding and Thomas Hofmann. Scalable non-equivariant 3d molecule generation via rotational alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267. PMLR, 2025. URL https://arxiv.org/abs/2506.10186.

P. Franco, N. Porta, J. D. Holliday, and P. Willett. The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation. *Journal of Cheminformatics*, 6:5, 2014.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1eWbxStPH.

Niklas W. A. Gebauer, Michael Gastegger, Stefaan S. P. Hessmann, Klaus-Robert Müller, and Kristof T. Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *arXiv preprint arXiv:2109.04824*, 2021. doi: 10.48550/arXiv.2109.04824.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. URL https://arxiv.org/abs/1912.01603.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning (ICML)*, pages 8867–8887. PMLR, 2022. URL https://arxiv.org/abs/2203.17003.

Ilia Igashov, Hannes Stärk, Clément Vignac, Víctor Garcia Satorras, Pascal Frossard, Max Welling, Michael M. Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design. In *NeurIPS AI for Science Workshop*, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin ídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Greg Landrum. Rdkit: Open-source cheminformatics software, 2016. URL https://www.rdkit.org. Accessed: 2026-06-10.

Y. C. Martin, J. L. Kofron, and L. M. Traphagen. Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358, 2002.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2022. URL https://arxiv.org/abs/2108.01073.

Minoru Nakata and Takahiro Maeda. Pubchemqc b3lyp/6-31g*//pm6 data set: The electronic structures of 86 million molecules using b3lyp/6-31g* calculations. *Journal of Chemical Information and Modeling*, 63(18): 5734–5754, 2023. doi: 10.1021/acs.jcim.3c00899. URL https://doi.org/10.1021/acs.jcim.3c00899.

Minoru Nakata and Tomohiro Shimazaki. Pubchemqc project: A large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling*, 57(6):1300–1308, 2017. doi: 10.1021/acs.jcim.7b00083. URL https://doi.org/10.1021/acs.jcim.7b00083.

Minoru Nakata, Tomohiro Shimazaki, Masahiro Hashimoto, and Takahiro Maeda. Pubchemqc pm6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *Journal of Chemical Information and Modeling*, 60(12):5891–5899, 2020. doi: 10.1021/acs.jcim.0c00740. URL https://doi.org/10.1021/acs.jcim.0c00740.

Ewa M. Nowara, Pedro O. Pinheiro, Sai Pooja Mahajan, Omar Mahmood, Andrew Martin Watkins, Saeed Saremi, and Michael Maser. Nebula: Neural empirical bayes under latent representations for efficient and controllable design of molecular libraries. *arXiv preprint arXiv:2407.03428*, 2024. doi: 10.48550/arXiv.2407.03428.

Ewa M. Nowara, Joshua Rackers, Patricia Suriana, Pan Kessel, Max Shen, Andrew Martin Watkins, and Michael Maser. Do we need equivariant models for molecule generation? *arXiv preprint arXiv:2507.09753*, 2025. doi: 10.48550/arXiv.2507.09753.

Pedro O. Pinheiro, Joshua Rackers, Joseph Kleinhenz, Michael Maser, Omar Mahmood, Andrew M. Watkins, Stephen Ra, Vishnu Sresht, and Saeed Saremi. 3d molecule generation by denoising voxel grids. *arXiv preprint arXiv:2306.07473*, 2023. doi: 10.48550/arXiv.2306.07473.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Learning a continuous representation of 3d molecular structures with deep generative models. *arXiv preprint arXiv:2010.08687*, 2020. URL https://arxiv.org/abs/2010.08687.

S. Riniker and G. A. Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5:26, 2013.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Saeed Saremi and Aapo Hyvärinen. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181): 1–23, 2019. URL https://jmlr.org/papers/v20/19-216.html.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Liò, Carla Gomes, Max Welling, et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022. doi: 10.48550/arXiv.2210.13695.

Ken Shoemake. Uniform random rotations. In *Graphics Gems III*, pages 124–132. Academic Press, 1992. doi: 10.1016/b978-0-08-050755-2.50036-1.

Gregor N. C. Simm, Robert Pinsler, Gábor Csányi, and José Miguel Hernández-Lobato. Symmetry-aware actor-critic for 3d molecular design. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=jEYKjPE1xYN.

Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, volume 11139 of *Lecture Notes in Computer Science*, pages 412–422. Springer, 2018. doi: 10.1007/978-3-030-01418-6_41.

Miha Skalic, José Jiménez, Davide Sabbadin, and Gianni De Fabritiis. Shape-based generative modeling for de novo drug design. *Journal of Chemical Information and Modeling*, 59(3):1205–1214, 2019.

S. Steshin. Lo-Hi: Practical ML drug discovery benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. In *NeurIPS*, 2018. URL https://arxiv.org/abs/1802.08219.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1711.00937.

Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. *arXiv preprint arXiv:2302.09048*, 2023. doi: 10.48550/arXiv.2302.09048.

Lvwei Wang, Rong Bai, Xiaoxuan Shi, Wei Zhang, Yinuo Cui, Xiaoman Wang, Cheng Wang, Haoyu Chang, Yingsheng Zhang, Jielong Zhou, et al. A pocket-based 3d molecule generative model fueled by experimental electron density. *Scientific Reports*, 12:15100, 2022.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nora Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Prashanth Venkatesh, Isaac Sappington, Susana V. Torres, Alston Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2203.02923.

A Technical appendices and supplementary material

A.1 Training and Architecture Details

VQ-VAE Compression Model. We use a VQ-VAE 3D autoencoder with no skip connections as the compression model. The architecture uses 4 resolution levels with channel multipliers $(1, 2, 2, 4)$, a base channel width of 32, self-attention on the lowest two resolutions, 2 residual blocks per level, GroupNorm with 16 groups, dropout of 0.1, and SiLU activations. The latent dimension is 1024 with 256 codebook embeddings of dimension 1024 and a commitment cost of 0.25. Input voxel grids are 64^3 with 8 element channels (C, H, O, N, F, S, Cl, Br) at 0.25 Åresolution. We train with batch size 32, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$), learning rate 10^{-5}, weight decay 10^{-2}, and EMA decay 0.999. Data augmentation consists of random $\mathrm{SO}(3)$ rotations and translations ($\delta = 0.5$). Training proceeds until mean IoU between input and reconstructed voxels exceeds 0.90.

Latent Denoising World Model. The latent denoiser is a 213M-parameter 3D U-Net operating on the $8^3 \times 1024$ latent space. It uses 96 base channels, channel multipliers $(1, 2, 2, 4)$, 2 blocks per level, GroupNorm with 32 groups, self-attention (with QK-norm) on the lowest two resolutions, and spatial-preserving down/upsampling to maintain the 8^3 spatial dimensions throughout. Skip connections concatenate encoder features to the decoder. Rotation conditioning is provided via Feature-wise Linear Modulation (FiLM): the 3×3 rotation matrix (flattened to 9 values) is embedded through a 5-layer MLP into a 3072-dimensional conditioning vector, which modulates each residual block's features via learned scale and shift parameters (unbounded linear FiLM, zero-initialized). We train with batch size 32, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$), learning rate 10^{-5}, weight decay 10^{-2},

cosine LR schedule, gradient clipping at 3.0, EMA decay 0.999, and noise level $\sigma = 1.8$. Training includes 15% identity rotation samples for stability. Training uses 100K subsampled molecules per epoch with DDP across 6 GPUs. For generation, we use Walk-Jump Sampling (WJS) with step size $\delta = 0.25$, friction 1.0, and Lipschitz constant 1.0. Training of the world model proceeds for 1500 epochs.

A.2 Generation with Walk-Jump Sampling

Following NEBULA (Nowara et al., 2024), we adopt Walk-Jump Sampling (WJS) (Saremi and Hyvärinen, 2019) as our generation procedure, operating entirely in the learned latent space. Unlike diffusion models that require iterative multi-step denoising with carefully tuned noise schedules, WJS achieves sample generation through a single noise level and one-shot denoising, making it substantially more efficient for molecular library design.

The sampling procedure consists of two alternating phases. In the *walk* phase, we draw samples from the noise-smoothed latent distribution via underdamped Langevin MCMC (Cheng et al., 2018) over k steps. This stochastic dynamics is governed by friction γ, inverse Lipschitz constant u, and the score function $\nabla \log p_\sigma(\mathbf{z})$, which is implicitly parameterized by our trained denoiser through Tweedie's identity. In the *jump* phase, a single forward pass of the denoiser maps the noisy sample back onto the data manifold, yielding a clean latent embedding.

The denoised latents are unnormalized using the precomputed channel statistics (μ_c, σ_c) and passed through the frozen VQ-VAE decoder to reconstruct 3D voxel grids. Atomic coordinates are then extracted from the voxels via optimization-based peak finding (Pinheiro et al., 2023), yielding the final molecular structures.

A.3 Uniform SO(3) Rotation Augmentation

To achieve rotation equivariance, we augment the training data with random 3D rotations drawn uniformly from the special orthogonal group $\mathrm{SO}(3)$. Naive parameterizations (e.g., Euler angles) induce non-uniform densities over the rotation manifold, leading to biased representations near the poles. We instead adopt the quaternion-based method of Shoemake (Shoemake, 1992), which guarantees a Haar-uniform distribution over $\mathrm{SO}(3)$.

Given three independent samples $u_1 \sim \mathrm{Uniform}(0,1)$, $u_2 \sim \mathrm{Uniform}(0,2\pi)$, $u_3 \sim \mathrm{Uniform}(0,2\pi)$, we construct a unit quaternion $\mathbf{q} = (q_w, q_x, q_y, q_z)$ as:

$$q_w = \sqrt{1-u_1}\,\sin u_2, \quad q_x = \sqrt{1-u_1}\,\cos u_2, \tag{17}$$

$$q_y = \sqrt{u_1}\,\sin u_3, \qquad q_z = \sqrt{u_1}\,\cos u_3. \tag{18}$$

The corresponding rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ is obtained via the standard quaternion-to-matrix conversion:

$$\mathbf{R} = \begin{pmatrix} 1-2(q_y^2+q_z^2) & 2(q_xq_y - q_zq_w) & 2(q_xq_z + q_yq_w) \\ 2(q_xq_y + q_zq_w) & 1-2(q_x^2+q_z^2) & 2(q_yq_z - q_xq_w) \\ 2(q_xq_z - q_yq_w) & 2(q_yq_z + q_xq_w) & 1-2(q_x^2+q_y^2) \end{pmatrix}. \tag{19}$$

During training, each molecule's atomic coordinates $\mathbf{x} \in \mathbb{R}^{N\times 3}$ are transformed as $\mathbf{x}' = \mathbf{x}\,\mathbf{R}^\top$, and the flattened rotation matrix $\mathrm{vec}(\mathbf{R}) \in \mathbb{R}^9$ is provided to the denoiser as a conditioning signal via FiLM. To stabilize training, 15% of samples retain the identity rotation $\mathbf{R} = \mathbf{I}_3$. At inference, conditioning on $\mathbf{I}_3$ generates molecules in canonical orientation, while conditioning on arbitrary $\mathbf{R}$ enables pose-controlled generation.

A.4 Quantification of the Error Budget

This appendix details the measurement of the constants reported in Sec. 3.5.

Protocol. All quantities are computed with the trained checkpoints (world model and frozen VQ-VAE) in inference mode, at the configuration used throughout the paper: $h = 0.25$ Å, 64^3 voxel grid, 8^3 latent grid ($h' = 2$ Å), $\sigma = 1.8$. Molecules are drawn from the held-out test split and passed through the same training-matched filter used during training, so that no evaluation is performed

on inputs the model never saw; this retains $13{,}055$ of $13{,}055$ candidates for the drugs split. Unless stated otherwise we use $10{,}000$ molecules $\times\, 8$ rotations drawn from μ_{Haar} by Shoemake's algorithm, giving $80{,}000$ (molecule, rotation) pairs; reported uncertainties are standard errors of the mean over those pairs. The physical origin and the map between physical and array axes are calibrated empirically by voxelizing single probe atoms rather than assumed, and the calibration is asserted to be a signed permutation (residual 0.0000 at every resolution tested).

Measured quantities. Writing $z = \Phi(\mathbf{x})$ and $\|\cdot\|_{\mathrm{rel}}$ for the relative ℓ_2 error, we report

- $\delta = \sqrt{\mathcal{L}(\theta^*)}$, the root-mean-square training residual, evaluated on the test split under the same objective equation 5 used for training;
- $\varepsilon_h = \big|\, \|\widetilde{T}_R(z)\|/\|z\| - 1 \,\big|$, the isometry defect of the induced latent action, which is the constant entering Eqs. equation 14–equation 16;
- the equivariance residual, the left-hand side of Eq. equation 8, obtained by evaluating the model on the rotated and the canonical branch and comparing the two outputs;
- the no-model baseline $\|\Phi(T_R\mathbf{x}) - \Phi(\mathbf{x})\|_{\mathrm{rel}}$, which is the residual an untrained identity map would incur.

Each is averaged over the same $80{,}000$ pairs.

Octahedral control. For R in the octahedral rotation group O the cubic lattice maps to itself, so the discretization contribution to the budget vanishes identically. Re-measuring the equivariance residual over O therefore isolates the part of the error that is *not* attributable to the grid. We verify independently that the lattice action on O is an exact permutation: the induced operators have singular values in $[0.999998, 1.000000]$ with exactly one non-zero entry per row. The residual changes from 0.1607 (Haar) to 0.1589 (O), so lattice effects account for roughly 1% of the budget and δ for the remainder.

Grid-resolution sweep. To confirm the $\mathcal{O}(h)$ rate assumed in Eq. equation 7 we hold the physical box fixed and vary the grid spacing, comparing the voxelization of the rotated molecule against the best trilinear reconstruction from the unrotated one. Table 4 reports the relative error and the empirical Lipschitz constant L of the density at each resolution.

Table 4: Voxel-level reconstruction error versus grid spacing, physical extent held fixed. The log–log slope is 1.55, between the $\mathcal{O}(h)$ Lipschitz rate and the $\mathcal{O}(h^2)$ C^2 rate, as expected for a density whose atomic width (0.5 Å) is comparable to h. The octahedral column is the lattice-preserving control.

h (Å)	grid	rel. error (Haar)	rel. error (O)	L (density)
1.000	16^3	1.0411	1.1817	0.462
0.500	32^3	0.3224	0.2698	0.972
0.250	64^3	0.1028	0.0586	1.531
0.125	128^3	0.0404	0.0112	1.758
log–log slope		1.551	—	—

A slope above 1 means the error vanishes *faster* than the analysis requires, so Eq. equation 7 is satisfied and conservative at the resolution we use.

Norm convention. Two conventions appear naturally in these measurements and must not be mixed. The objective $\mathcal{L}$ is an expectation of $\|\cdot\|^2$ over a latent of dimension $n = C{\cdot}|\Lambda'_{h'}| = 524{,}288$, and is naturally reported as a *per-element* root mean square, $\|v\|_{\mathrm{rms}} = \|v\|_2/\sqrt{n}$; the equivariance residual and the sweep of Table 4 are naturally reported as *relative* errors $\|a-b\|_2/\|b\|_2$. The two differ by the scale of the reference vector, and a third (vector-ℓ_2) convention differs from the first by $\sqrt{n} = 724.1$, so a mismatch can be large. We therefore measure the scale of the normalised latent directly: over 300 molecules, $\|z\|_{\mathrm{rms}} = 0.7916 \pm 0.0028$ (p05 0.712, p95 0.878). Since $\mathcal{L}(\theta^*)$ gives $\delta_{\mathrm{rms}} = 0.1549$, the same residual expressed relatively is $\delta = \delta_{\mathrm{rms}}/\|z\|_{\mathrm{rms}} = 0.1957$, which is the

value quoted in Table 1. Reporting the whole budget in either convention gives the same verdict, as it must:

convention	certificate $2\delta(1+\varepsilon_h)$	measured residual	margin
relative, $\|a-b\|_2/\|b\|_2$	0.3984	0.1607	2.48×
per-element, $\|\cdot\|_2/\sqrt{n}$	0.3154	0.1272	2.48×

The bound is therefore satisfied with a factor of 2.5 to spare, and the margin is a property of the measurement rather than of the convention chosen to express it.

Scope of ε_h. Eq. equation 15 bounds only the voxel-level contribution to ε_h. The measured value also absorbs any non-equivariance of the learned encoder, which is not a function of h and which we therefore report rather than bound: the VQ-VAE is a conventional convolutional encoder and is not constrained to commute with rotations. This is why ε_h is quoted as a measurement in Table 1. It does not affect the validity of Theorem 1, whose hypotheses concern the latent action and the convergence of the objective, but it does mean the $\mathcal{O}(h)$ statement should be read as describing the grid contribution alone. An encoder with steerable kernels would remove the remaining term; we leave this to future work.

A.5 Ablation Study: Withholding the Rotation Hint

This appendix expands the ablation summarised in the main text. The question it addresses is whether the world-model formulation — predicting the action that undoes a known rotation — is doing work that rotation augmentation alone would achieve.

Variants and controls. Two variants withhold R from the network:

- ZERO-COND: the rotation embedder and every FiLM layer are retained and continue to receive gradients, but the 9-dimensional flattened rotation matrix is replaced by a constant zero vector. The parameter count is unchanged at 213.31M, so any degradation cannot be attributed to reduced capacity.
- NO-FILM: the rotation embedder and all FiLM layers are removed, leaving a 102.35M-parameter backbone. This variant guards against the objection that zeroed FiLM layers might still carry information through their learned biases.

Two deliberate controls make the outcome difficult to attribute to training pathology. First, an *identity-input curve* is logged every epoch: the same weights are evaluated on un-rotated inputs, so a model that denoises well there but fails on rotated inputs has a failure localised to the rotation term. Second, the learning rate is held *constant* at 10^{-6} with no schedule, so a flat loss cannot be attributed to an annealed step size. Everything else — data pipeline, frozen VQ-VAE weights, latent normalisation statistics, $\sigma = 1.8$, batch size, optimiser and the 30-epoch budget — is copied verbatim from the main run.

Table 5: Final values after 30 epochs. The two variants land in the same place despite a 2.1× difference in parameter count, which rules out both capacity and residual information flow through the zeroed FiLM layers. The frozen VQ-VAE round-trip IoU is 0.906.

Final value	ZERO-COND	NO-FILM
validation loss, rotated input	0.4943	0.4980
predict-the-mean baseline	0.5747	0.5738
validation loss, identity input	0.0315	0.0319
reconstruction IoU, rotated	0.0058	0.0058
reconstruction IoU, identity	0.7598	0.7746
rotation consistency (spread over K poses)	0.0044	0.0071
prediction std (target 1.0)	0.3018	0.3027
parameters	213.31M	102.35M

Why the failure is informational. When R is not observable, the minimiser of the squared error is the conditional mean $\mathbb{E}[z_{\text{can}} \mid z_{\text{rot}}]$, an average over the $SO(3)$ orbit of the molecule. A network at that optimum returns one pose-independent answer for every rotation of the same input, which is precisely what the diagnostics record: the spread of predictions across K poses of a single molecule falls to 0.0044 while the corresponding inputs remain 1.43 apart (Fig. 3, right), and the prediction standard deviation settles at 0.302 against a target of 1.0 (left). The objective is not merely harder without the hint; it has a different and degenerate optimum.

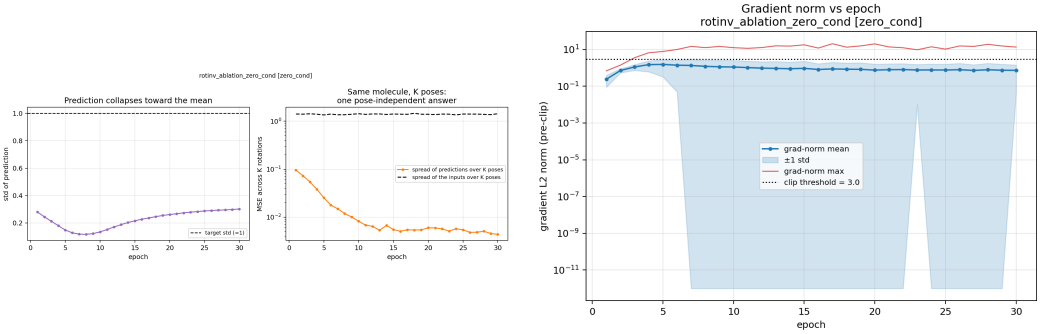

Figure 3: **Left:** prediction collapse. Prediction standard deviation never approaches its target of 1.0, and the spread of outputs over K poses of one molecule falls by two orders of magnitude while the inputs stay 1.43 apart — one answer for the whole orbit. **Right:** gradient norms. Means rise from 0.243 to 0.734 with maxima near 10 and no non-finite steps, and the learning rate is never decayed. Nothing diverged: the flat loss in Fig. 2 is not an optimisation artefact.

The identity fraction. The augmentation schedule reserves 15% of each batch for canonical (unrotated) examples. Setting this fraction to zero removes the only sub-task whose target is unambiguous, and the collapse deepens: identity-input IoU falls from 0.7746 to 0.0023, the ratio of task loss to the predict-the-mean baseline reaches 1.016, and prediction standard deviation falls to 0.117 (Fig. 4). Gradient norms again remain well-behaved, rising gently from 0.686 to 0.899. The identity fraction therefore acts as an anchor that keeps the regression target well-posed during training rather than as a shortcut that leaks the answer — a distinction worth drawing explicitly, since a reader might otherwise read the canonical examples as making the task artificially easy.

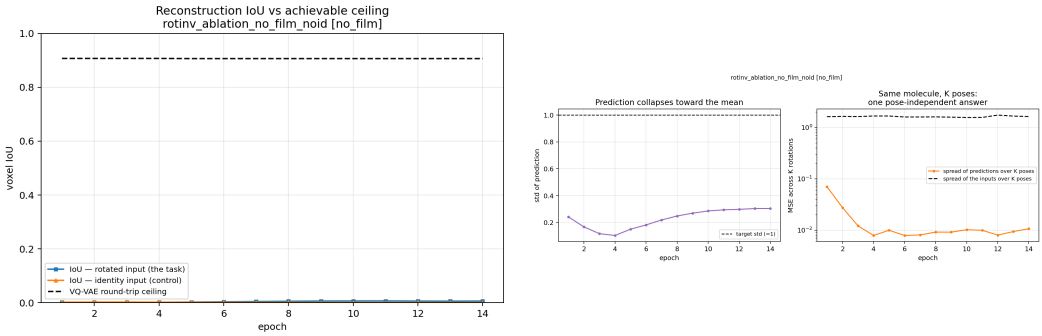

Figure 4: Removing the 15% identity fraction (NO-FILM, identity fraction 0). **Left:** both IoU curves now sit on the floor — without canonical examples the model no longer learns even the identity sub-task. **Right:** the corresponding collapse diagnostics, with prediction standard deviation at 0.117 against a target of 1.0.

Scope. The ablation establishes that withholding R makes the task unlearnable at matched capacity, with the denoising ability of the same weights retained as a control. It does not, and cannot, quantify how far a *differently formulated* augmentation-only model would get: removing R changes the population optimum rather than simply removing a hint, so the comparison is between a well-posed and an ill-posed objective rather than between two competing designs. A model trained to

predict the canonical latent without ever conditioning on pose — for instance a denoiser trained only on canonical molecules — is a different and weaker baseline, and is reported separately among the non-equivariant baselines in Table 2.

A.6 Evaluation Metrics

We assess generation quality using the metrics established by Vignac et al. (2023). **Atom stability** measures the fraction of atoms whose bond count equals their expected valence, and **molecule stability** requires all atoms within a molecule to be individually stable (Hoogeboom et al., 2022). **Validity** is the percentage of generated structures that pass RDKit (Landrum, 2016) sanitization checks. To compare distributional fidelity against the reference dataset, we report: **Valency** W_1, the 1-Wasserstein distance between valence distributions; **Atoms TV** and **Bonds TV**, the total variation distances between atom-type and bond-type distributions, respectively; and **Bond Length** W_1 and **Bond Angle** W_1, the 1-Wasserstein distances between the bond length and bond angle distributions of generated and reference molecules. For seeded (library) generation, we additionally report the **Tanimoto similarity** between each generated molecule and its seed, computed using Morgan fingerprints (radius 2, 2048 bits) (Rogers and Hahn, 2010), averaged over all valid seed–generation pairs.

A.7 Seed-Conditioned Drug Candidate Generation

To demonstrate the practical utility of FLDWM beyond benchmark metrics, we performed seed-conditioned generation against five FDA-approved drugs that collectively span four therapeutic areas with documented unmet needs: **Bedaquiline** (multi-drug-resistant tuberculosis; QTc prolongation in $\sim$11% of patients and emerging *atpE*/*Rv0678* resistance), **Osimertinib** (*EGFR* T790M NSCLC; $\sim$\$15K/mo list price and acquired C797S resistance), **Sotorasib** (*KRAS* G12C NSCLC; objective response plateau at $\sim$37%), **Imatinib** (BCR–ABL CML; later-line TKIs exceed \$100K/yr and the T315I gatekeeper mutation persists), and **Ivacaftor** (*CFTR* G551D cystic fibrosis; $\sim$\$311K/patient/yr, not cost-effective at standard willingness-to-pay thresholds). For each seed, the molecule was voxelised, encoded by the trained VQ-VAE into the $8^3 \times 256$ latent grid, perturbed with σ=1.8 Gaussian noise, denoised by the FiLM-conditioned world model under Walk–Jump Sampling (WJS) for $n \in \{10, 20, 50, 100, 200\}$ steps, and decoded back to an atomic structure. Candidates were retained only if they satisfied RDKit validity, MiDi-style molecule stability, single-component connectivity, and ECFP4 (Morgan r=2, 2048 bit) Tanimoto similarity $T_c \geq 0.3$ to the seed. A total of 125 candidates were generated across the five drugs and five WJS settings.

Threshold selection. We adopt the *Lead Optimization* window $T_c \in [0.4, 0.7]$ on ECFP4 as the operational range for "promising" analogs, justified by three converging lines of evidence. *(i) Activity conservation:* on ChEMBL, $\sim$ 37.5% of known active–active pairs against a common target exceed $T_c \geq 0.3$ on ECFP4 versus only $\sim$ 0.03% for active–random pairs Martin et al. (2002). *(ii) Regulatory recognition:* the EMA/FDA orphan-drug study by Franco *et al.* mapped the binary similarity judgments of 143 regulatory experts against algorithmic Tanimoto scores and identified ECFP4 $T_c \approx 0.4$ as the *tipping point* above which experts deem two molecules to share "the same principal structural features" Franco et al. (2014); this same boundary is formalised in the NeurIPS 2023 Lo–Hi benchmark separating Lead Optimisation ($T_c > 0.4$) from Hit Identification ($T_c \leq 0.4$) Steshin (2023). *(iii) IP defensibility:* above $T_c \approx 0.7$ generated molecules become near-trivial substitutions of the seed, are classified as redundant "me-too" analogs, and likely infringe the originator's composition-of-matter patentdefeating the affordability rationale of the analog programme. We therefore explicitly reject the historical $T_c \geq 0.85$ heuristic, which retrospective HTS studies showed to deliver only $\sim$30% activity-sharing probability and which cannot be transposed across fingerprint algorithms Martin et al. (2002); Riniker and Landrum (2013). ECFP4 was preferred over MACCS keys and over modern transformer embeddings (ChemBERTa, SELFormer) on the basis of the Riniker–Landrum benchmark and the recent ECFP4-vs-SELFIES retrieval study, in which ECFP4 + Tanimoto attained the lowest mean graph-edit distance on analog retrieval Riniker and Landrum (2013); smi (2024).

Results. Of the 125 generated candidates, 49 fell in the lead-optimisation window; after deduplication, the candidate closest to the mid-range T_c=0.55 was retained per drug as the headline output (Table 6). All five drugs yielded a chemically distinct, IP-defensible analog with $T_c \in [0.50, 0.64]$,

consistent with the ECFP4 lead-optimisation regime. Visual side-by-side comparisons of seed and analog 2D structures are provided in Supplementary Figures S1–S5.

Table 6: Lead-optimisation analogs generated by FLDWM for five FDA-approved seed drugs. ECFP4 Tanimoto values are computed against the seed using RDKit `GetMorganFingerprintAsBitVect`(r=2, n=2048).

Seed drug	**Indication**	**WJS steps**	**ECFP4** T_c
Osimertinib	NSCLC (*EGFR* T790M)	200	0.6322
Imatinib	CML (BCR–ABL)	100	0.5385
Sotorasib	NSCLC (*KRAS* G12C)	100	0.5294
Ivacaftor	Cystic fibrosis (*CFTR* G551D)	20	0.5238
Bedaquiline	Multi-drug-resistant tuberculosis	50	0.5062

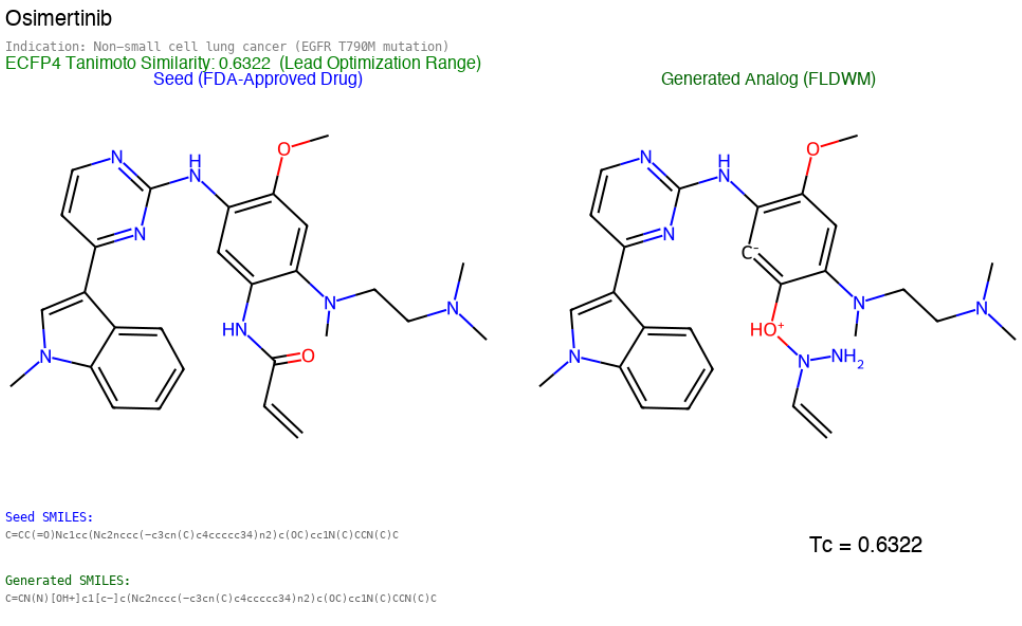

Figure 5: FLDWM-generated lead-optimization analog of **Osimertinib** (NSCLC, EGFR T790M). ECFP4 Tanimoto similarity to the seed: $T_c = 0.6322$ (WJS 200 steps).

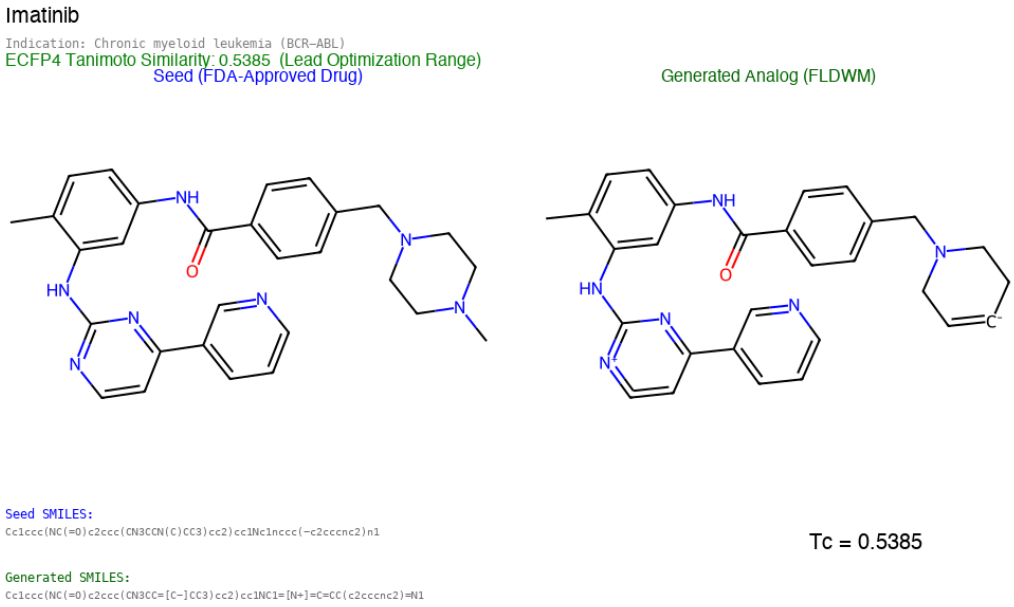

Figure 6: FLDWM-generated lead-optimization analog of **Imatinib** (CML, BCR–ABL). ECFP4 Tanimoto similarity to the seed: $T_c = 0.5385$ (WJS 100 steps).

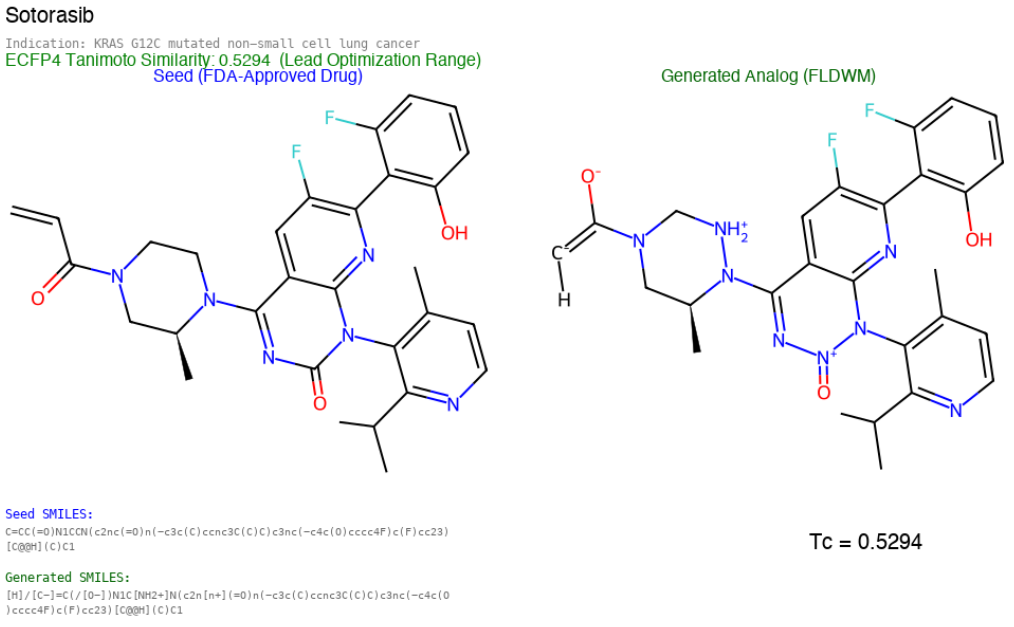

Figure 7: FLDWM-generated lead-optimization analog of **Sotorasib** (NSCLC, KRAS G12C). ECFP4 Tanimoto similarity to the seed: $T_c = 0.5294$ (WJS 100 steps).

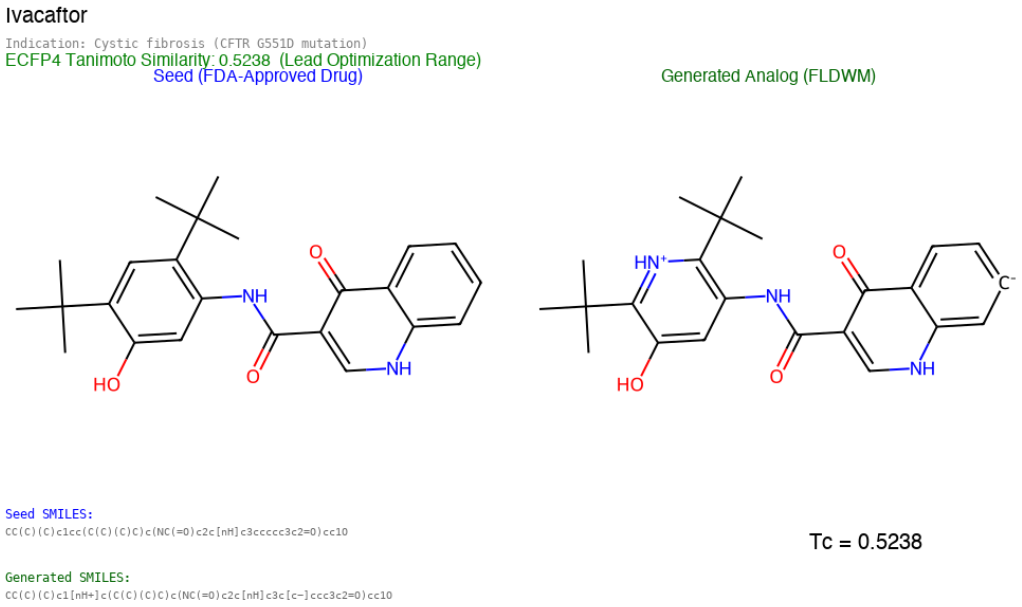

Figure 8: FLDWM-generated lead-optimization analog of **Ivacaftor** (cystic fibrosis, CFTR G551D). ECFP4 Tanimoto similarity to the seed: $T_c = 0.5238$ (WJS 20 steps).

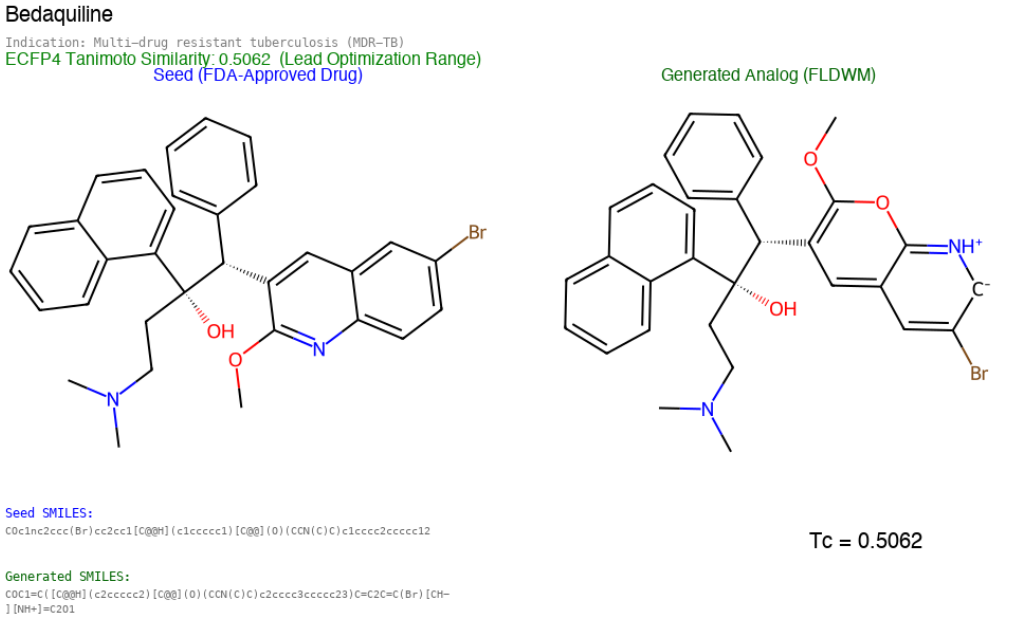

Figure 9: FLDWM-generated lead-optimization analog of **Bedaquiline** (multi-drug-resistant tuberculosis). ECFP4 Tanimoto similarity to the seed: $T_c = 0.5062$ (WJS 50 steps).