Do we need equivariant models for molecule generation?

Ewa M. Nowara [1] **Joshua Rackers** [2] **Patricia Suriana** [1]
Pan Kessel [3] **Max Shen** [1] **Andrew Martin Watkins** [1] **Michael Maser** [1]

Abstract

Deep generative models are increasingly used for molecular discovery, with most recent approaches relying on equivariant graph neural networks (GNNs) under the assumption that explicit equivariance is essential for generating high-quality 3D molecules. However, these models are complex, difficult to train, and scale poorly.

We investigate whether non-equivariant convolutional neural networks (CNNs) trained with rotation augmentations can learn equivariance and match the performance of equivariant models. We derive a loss decomposition that separates prediction error from equivariance error, and evaluate how model size, dataset size, and training duration affect performance across denoising, molecule generation, and property prediction. To our knowledge, this is the first study to analyze learned equivariance in generative tasks.

1. Introduction

Generative models have transformed fields like computer vision and natural language processnig, and are increasingly applied in drug discovery to design novel molecules. Many recent approaches rely on equivariant deep learning, particularly equivariant graph neural networks (GNNs) (Hoogeboom et al., 2022; Gebauer et al., 2019; Vignac et al., 2023; Bao et al., 2022; Huang et al., 2024; Xu et al., 2023), based on the assumption that equivariance to 3D rotations and translations is essential for high-quality molecular structures. However, enforcing equivariance through architectural design introduces practical limitations: these models are harder to scale, slower to train, and more difficult to implement.

An alternative is to use data augmentation, applying random rotations and translations during training, which encourages learned equivariance without architectural constraints. This allows the use of more flexible architectures such as 3D convolutional neural networks (CNNs). Recent CNN-based models (Pinheiro et al., 2023; 2024; Nowara et al., 2024) have outperformed equivariant GNNs, despite lacking explicit equivariance.

We investigate whether CNNs trained with data augmentation learn an equivariant behavior, and how model size, dataset size, and training time influence it. We study this across three tasks: molecule reconstruction from noise (denoising), molecule generation, and property prediction using the public GEOM-Drugs dataset (Axelrod & Gomez-Bombarelli, 2022).

While equivariance has been studied extensively in supervised tasks (e.g., classification, segmentation (Gerken & Kessel, 2024; Quiroga et al., 2020; Gerken et al., 2022)), its role in generative modeling remains underexplored. To our knowledge, this is the first systematic evaluation of learned equivariance in molecular generative models. We focus on *seeded* generation (starting from a given lead compound), where equivariance is especially critical. If outputs vary under input rotations, the model becomes unreliable and computationally inefficient.

We find that CNNs trained with rotation augmentation easily learn equivariance during denoising—even with small models, limited data (as little as 10%), and few training epochs. Property prediction shows similar robustness. However, in generative tasks, only larger models trained on more data produce consistent outputs across rotated seeds. Smaller models diverge, losing generative equivariance.

Interestingly, latent embeddings of rotated molecules differ significantly, even when reconstructions and predictions are nearly identical, suggesting that CNNs learn redundant parallel representations instead of recognizing rotated inputs as equivalent. This inefficiency may explain their higher parameter demands compared to equivariant GNNs. We hypothesize that auxiliary training objectives used to align latent embeddings across rotations could improve robustness without requiring large models or datasets.

[1]Prescient Design, Genentech, South San Francisco, USA [2]Achira [3]Prescient Design, Roche, Basel, Switzerland. Correspondence to: Ewa M. Nowara <nowara.ewa@gene.com>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

2. Related Work

2.1. Equivariance in Machine Learning

Several works compare data augmentation and equivariant architectures. See Bronstein et al. (2021) for an overview of geometric deep learning and Fei & Deng (2024) for a survey of equivariance in 3D machine learning. Kvinge et al. (2022) introduced metrics for evaluating invariance and equivariance in both approaches. Gerken & Kessel (2024) showed that equivariance can emerge from data augmentation in predictive tasks, and Quiroga et al. (2020) found no advantage in using equivariant CNNs for image classification. However, Gerken et al. (2022) showed that equivariant models significantly outperform augmented CNNs on spherical image segmentation, which they argue is a more challenging setting. They also found that larger models and datasets help, but are not sufficient to match equivariant architectures. Our work is the first to study how learned equivariance affects the distribution of generated samples in generative models.

2.2. Molecule Generation

3D molecular generation improves over 2D representations by capturing spatial structure. Existing models typically use either point cloud representations with explicitly equivariant GNNs, or voxel-based CNNs with data augmentation.

GSchNet (Gebauer et al., 2019) introduced point-set autoregressive generation. EDM (Hoogeboom et al., 2022) used SE(3)-equivariant diffusion to denoise point clouds from Gaussian noise. MiDi (Vignac et al., 2023) jointly generated molecular graphs and 3D conformations.

Voxel-based models use CNNs for generation. Early works (Skalic et al., 2019; Ragoza et al., 2020) applied VAEs (Kingma & Welling, 2014) to voxelized inputs. VoxMol (Pinheiro et al., 2023) used walk-jump sampling (WJS) (Saremi & Hyvärinen, 2019), a diffusion-inspired process, for efficient generation. Nebula (Nowara et al., 2024) trained a latent generative model on voxel embeddings to improve generalization. Voxel models have also been extended for structure-based generation conditioned on a protein pocket (Pinheiro et al., 2024).

3. Background

3.1. Equivariance

We begin by defining equivariance and approximate equivariance that arises in models trained with data augmentation.

Definition 3.1 (Equivariant Functions)**.** A function f is equivariant to a transformation R if applying R to the input x corresponds to applying the transformation R to the output, such that

$$f(R(x)) = R(f(x)) \tag{1}$$

In this work, we focus on SE(3) equivariance, encompassing both rotations and translations. SE(3) equivariance excludes reflections, which can alter molecular properties, so we typically do not want models to be equivariant to them. Since convolutional neural networks (CNNs) are translationally equivariant by design, we primarily study rotational equivariance.

Theorem 3.2 (Loss decomposition for quantifying learned equivariance)**.** *For a function $f : \mathbb{R}^D \to \mathbb{R}^D$, a loss function $l : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, a dataset X with labels Y, and a class of transformations R,*

$$\mathbb{E}_{x,R}[l(f(R(x)), R(y))] \approx \underbrace{\mathbb{E}_x\left[l\left(\mathbb{E}_R[R^{-1}f(R(x))], y\right)\right]}_{\text{prediction error}} + \underbrace{\frac{1}{2}\mathbb{E}_x\left[\text{tr}\left(H_l(\mu, y)\,\text{Var}_R(R^{-1}f(R(x)))\right)\right]}_{\text{equivariance error}} \tag{2}$$

where $H_l(\mu, y)$ is the Hessian of the loss function l given terms $\mu = \mathbb{E}_R[R^{-1}f(R(x))]$ and y, and Var is a $D \times D$ covariance matrix.

Moreover, if l is MSE loss, then the relation holds exactly, and simplifies to:

$$\mathbb{E}_{x,R}\left[l(f(R(x)), R(y))\right] = \underbrace{\mathbb{E}_x\left[l\left(\mathbb{E}_R\left[R^{-1}f(R(x))\right], y\right)\right]}_{\text{prediction error}} + \underbrace{\frac{1}{D}\mathbb{E}_x\left[\sum_{i=1}^{D}\text{Var}_R\left(\left[R^{-1}f(R(x))\right]_i\right)\right]}_{\text{equivariance error}} \tag{3}$$

Proof. This follows by a second-order Taylor expansion around μ. □

Definition 3.3 (Approximate Equivariance)**.** A model trained with rotation augmentation can learn approximate equivariance, even if its architecture is not explicitly equivariant.

For a denoising model, motivated by theorem 3.2, we quantify rotational equivariance by comparing the reconstruction of a rotated input molecule $\hat{x}_n$ to the rotated reconstruction of the unrotated input $R(\hat{x}_0)$. An equivariant model should satisfy:

$$||R(\hat{x}_0) - \hat{x}_n||_2^2 < \epsilon \tag{4}$$

where ϵ is a small threshold corresponding to the reconstruction error floor, and squared L2-norm is related to the variance term in theorem 3.2.

3.2. Equivariance of the Denoising Process

Let x denote clean data (e.g., a molecule), and y the corresponding noisy input after adding isotropic Gaussian noise. The noisy data distribution is:

$$p(y) = \int p(x)p(y|x)dx \tag{5}$$

Assuming the data distribution and the noise process are both rotation-equivariant:

$$\begin{aligned} p(Rx) &= p(x) \\ p(Ry|Rx) &= p(y|x) \end{aligned} \tag{6}$$

Then, we can show that the smoothed noisy distribution $p(y)$ is also equivariant if smoothing and data density is also equivariant, by substituting $x = R\tilde{x}$.

$$\begin{aligned} p(Ry) &= \int p(x)\, p(Ry \mid x)\, dx \\ &= \int p(R\tilde{x})\, p(Ry \mid R\tilde{x})\, d\tilde{x} \quad (\text{where } x = R\tilde{x}) \\ &\text{(since } \det R = 1, \text{ we have } dx = d(R\tilde{x}) = d\tilde{x}) \\ &= \int p(\tilde{x})\, p(y \mid \tilde{x})\, d\tilde{x} \\ &= p(y) \end{aligned} \tag{7}$$

Let D be the denoising function. The denoised distribution is:

$$\hat{p}(x) = \int \delta(x - D(y))\, p(y)\, dy \tag{8}$$

If the denoiser is equivariant, i.e., $D(Ry) = RD(y)$, and $p(Ry) = p(y)$ as above, then, by substituting $y = R\tilde{y}$.

$$\begin{aligned} \hat{p}(x) &= \int \delta(Rx - D(y))\, p(y)\, dy \\ &= \int \delta(Rx - D(R\tilde{y}))\, p(R\tilde{y})\, d\tilde{y} \\ &= \int \delta(R(x - D(\tilde{y})))\, p(\tilde{y})\, d\tilde{y} \\ &= \hat{p}(x) \end{aligned} \tag{9}$$

Using the fact that $\delta(R(x - D(y))) = \delta(x - D(y))$ for any invertible matrix R with $\det R = 1$ which holds for all $R \in \mathrm{SO}(3)$ the delta function is invariant under such transformations. Thus, the denoised distribution is rotation-equivariant if both the noise and the denoiser are equivariant.

3.3. Molecule Generation

Representing Molecules as 3D Voxels. Following (Pinheiro et al., 2023), we represent each molecule as 3D voxels by drawing a continuous Gaussian density around each atom's atomic coordinates on a voxel grid and represent each atom type as a separate channel, resulting in a 4D tensor of $[c \times l \times l \times l]$ for each molecule, where c is the number of atom types and l is the length of the voxel grid edge. The voxel values are normalized to a 0 to 1 range. See Fig 1 for an example of a voxel representation of a molecule.

Generating Molecules with Walk Jump Sampling. We generate new molecules with a sampling process called *neural empirical* Bayes (Saremi & Hyvärinen, 2019). It is a two-step score-based sampling method, called *walk-jump sampling* (WJS). In the first "walk" step, we sample new noisy molecular voxel grid with multiple k walk steps of a Langevin Markov chain Monte Carlo (MCMC) sampling (Cheng et al., 2018) along a randomly initialized manifold (eq. 10). In the second "jump" step, we denoise the newly sampled noisy molecule grid with a forward pass of a pre-trained denoising autoencoder (DAE) model at an arbitrary step k (eq. 11). The DAE is trained to denoise voxelized molecules with random isotropic Gaussian noise added to their voxel grid with a mean squared error (MSE) loss between ground truth and reconstructed voxels. WJS is very similar to diffusion models but it is much faster as it only requires a single noising and denoising step (Pinheiro et al., 2023; Nowara et al., 2024; Pinheiro et al., 2024).

$$\begin{aligned} dv_t &= -\gamma v_t dt - u g_\theta(y_t)dt + (\sqrt{2\gamma u})dB_t, \\ dy_t &= v_t dt, \end{aligned} \tag{10}$$

where B_t is Brownian motion, γ is friction, u is inverse mass, and γ, u are hyperparameters, and g_θ is the learned denoising score function.

$$\hat{x} = g_\theta(y) \tag{11}$$

To obtain a molecular graph from the generated voxels, we localize the atomic coordinates as the center of each voxel using an optimization-based peak finding method (Pinheiro et al., 2023).

4. Experiments and Results

4.1. Reconstruction

Generative models are trained to denoise the input from Gaussian noise (Pinheiro et al., 2023; Nowara et al., 2024; Rombach et al., 2022). If a model is unable to properly reconstruct (denoise) the input, the generation quality will be poor. Therefore, first we study the quality of the molecule reconstruction and how it's affected by molecule rotations.

To evaluate whether generative models learn rotational equivariance, we measure reconstruction equivariance to rotations of the input molecule. By definition of equivariance in eq. 4, we compare the reconstructed molecule from a rotated input $\hat{x}_n$ to the rotated reconstruction of an

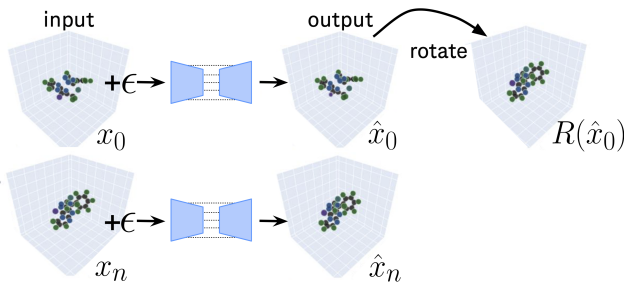

Figure 1. Experimental set up of reconstruction equivariance to rotations of voxelized molecules. We compute the reconstruction error between reconstructed molecules from a rotated input $\hat{x}_n$ to the rotated reconstruction of an unrotated input $R(\hat{x}_0)$.

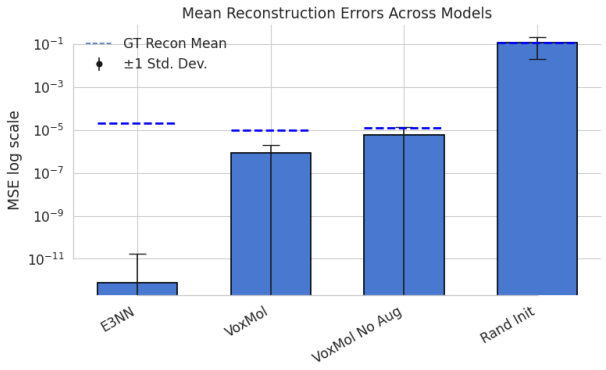

Figure 2. Rotational equivariance with different models compared to reconstruction upper bound computed between reconstructed and ground truth molecule (dotted blue lines).

unrotated input $R(\hat{x}_0)$. A model is equivariant if these outputs are identical or differ only by a small, bounded error. Figure 1 shows the experimental set up for reconstruction equivariance.

We compare VoxMol (Pinheiro et al., 2023) trained with rotation and translation augmentations and its equivariant version (E3NN). Due to prohibitively long training times of the E3NN, we were only able to train a smaller model with 0.5M parameters compared to 111M in VoxMol. The training cost of E3NN per epoch was 90 minutes compared to only 2 hours for the full size VoxMol which has 200 times more parameters.

Note that we use the same amount of training data for VoxMol trained with and without data augmentation. Rather than increasing the dataset size by presenting each molecule in multiple rotated forms, we apply a random rotation to each molecule once per training iteration. This preserves the effective training set size while introducing rotational variability. Detailed results with reconstruction errors for each rotation axis and angle are reported for each experiment in the Appendix A.3. See Appendix A.1 for architecture, training details, and dataset A.2.

Figure 2 shows that both equivariant E3NN and VoxMol trained with rotation augmentations achieve low equivariant reconstruction error. Notably, these errors are consistently lower than the ground truth reconstruction error (between the input and the output in 3D) which can be regarded as an upper bound of reconstruction capability of a model, indicating strong approximate equivariance even for models without explicitly equivariant architectures. While E3NN achieves a lower reconstruction error and lower equivariance error than VoxMol, it does not translate to better molecule generation quality, as will be demonstrated in the next section.

Training with No Augmentations. Interestingly, VoxMol trained without rotation augmentation also demonstrates a lower equivariant error than ground truth reconstruction error, suggesting that the 3D CNN architecture and the denoising task inherently capture some rotational structure from the sparse voxel inputs.

Random Initialization. To verify whether equivariance arises automatically from the architecture or from the sparsity of the voxel representation, we evaluate the reconstructions with a randomly initialized VoxMol with no training. However, we find that the untrained model does not exhibit reconstruction equivariance although it can be learned very quickly during training. This confirms that equivariance is not an inherent property of the architecture or the voxel inputs, but rather a learned behavior that emerges during training with rotation augmented data.

Effect of Model Size. We also evaluate VoxMol models with significantly fewer parameters 28 million (M) and 7 M, compared to the full 111 M model in Figure 3. Despite reduced reconstruction accuracy, both smaller models exhibit similarly lower equivariant reconstruction error compared to the upper bound. This suggests that model size has little impact on a model's ability to learn equivariance from rotation augmentation, even though it does affect the overall quality of reconstruction and consequently will affect the quality of the generations.

Effect of Training Set Size. Prior work has suggested that large datasets and long training times are needed for the learned equivariance to emerge (Gerken et al., 2022). Contrary to that belief, we find that VoxMol learns strong reconstruction equivariance even when trained on as little as 1% of the dataset. Figure 4 shows that the equivariant reconstruction error remains consistently low across models trained on 50% (550K), 25% (275K), 10% (110K), and 1% (11K) of the data (the full dataset contains 1.1M molecules), indicating that data augmentation alone is sufficient to induce equivariance even in low-data regimes.

Effect of Training Duration. Similarly, we assess models trained for 100 to 1,000 epochs (on the 10% training set

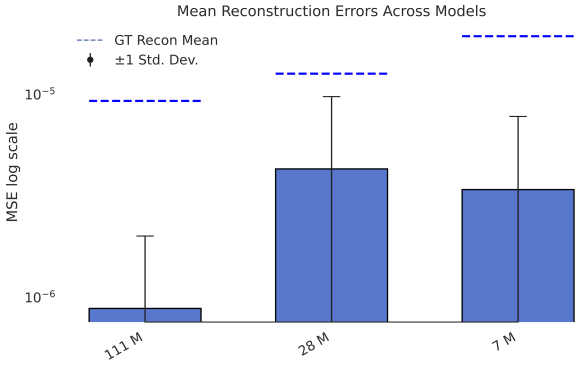

Figure 3. Effect of **model size** on reconstruction equivariance. Even much smaller models are able to learn equivariance during reconstruction.

Figure 4. Effect of **training set size**. Equivariance is already learned with limited data although the overall reconstruction errors increase with decreasing amount of training data.

subset) in Figure 5. Equivariance is learned early during training, with marginal improvements from longer training. While training for more epochs improves the overall reconstruction quality and consequently the generation performance, they do not meaningfully affect reconstruction equivariance.

Latent Embeddings. Since the model is easily able to learn reconstruction equivariance without large training set sizes, with few training epochs, and even without explicit data augmentation, we seek to understand what feature of the training process leads to learned equivariance. Previous works have analyzed the latent embeddings of models to study the extent of their equivariance (Kvinge et al., 2022). An equivariant model should embed an input object to the same latent embeddings regardless of its rotation.

We compute the cosine similarity between latent embeddings of the same molecule under different rotations and find that, except at 0°or 360°, VoxMol's embeddings differ significantly (see Figure 6). This indicates that the model does not recognize rotated molecules as the same object in

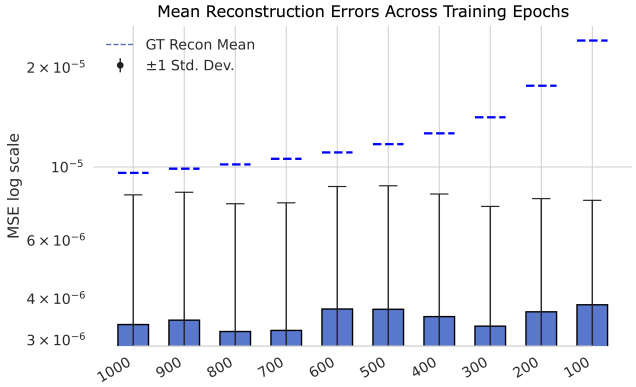

Figure 5. Effect of the **number of training epochs**. Equivariance is learned in very early epochs although the overall reconstruction errors are higher in earlier epochs.

latent space.

However, we do observe that VoxMol trained with rotation augmentations embeds the rotated molecules closer together compared to VoxMol trained with no augmentations. In contrast, the explicitly equivariant E3NN model produces identical latent embeddings for rotated inputs, as expected (shown in Appendix Figure 16).

These findings suggest that while VoxMol learns to reconstruct rotated inputs correctly, it does so by learning redundant latent representations. This inefficiency may require larger model capacity and could hinder tasks that rely on meaningful latent structure, such as property prediction or property guided generation (Kaufman et al., 2024; Xu et al., 2023; Huang et al., 2024).

4.2. Molecule Generation

Reconstruction equivariance does not necessarily imply equivariance during generation nor does it guarantee high-quality generations. To evaluate how rotations affect the distribution of generated molecules, we conduct a series of generation experiments to assess whether models maintain equivariance beyond denoising. We perform two controlled experiments for each model. In the first experiment, we initialize generation with the same seed lead molecule rotated by 0°, 90°, 180°, 270°, and 360°, testing how rotation impacts the generations. In the second experiment, we fix the seed molecule's orientation and generate the same number of molecules as in the first experiment using the same noising for reproducibility, to test how randomness from "Brownian motion" alone in WJS affects the generations compared to rotations of the seed molecule. Figure 7 shows the experimental set up for generation equivariance.

In both experiments, we run 5 chains for 100 denoising steps each, saving intermediate molecules every 10 steps, repeated for 100 seed ligands. This results in 5,000 generated

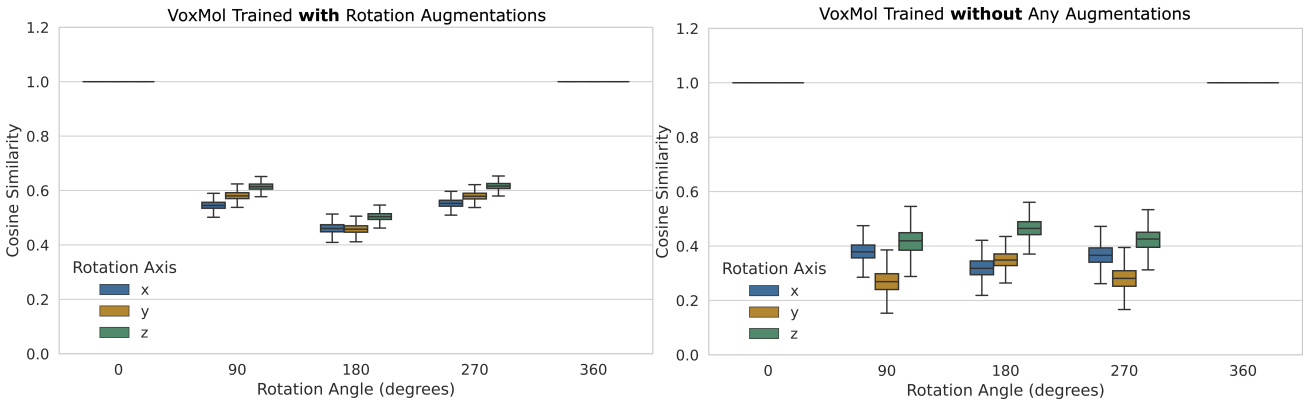

Figure 6. Cosine similarity between latent embeddings of the same molecules under different rotations. Latent embeddings are not rotation equivariant as rotated molecules are embedded differently. However, VoxMol trained with rotation augmentations embeds rotated molecules closer to each other compared to VoxMol trained with no augmentations, demostrating some learned equivariance in the latent space.

	stable %↑	valid %↑	unique %↑	atom TV↓	bond TV↓	bond ang W1↓
E3NN - Unrotated Input	41 (±16)	85 (±11)	95 (±6)	0.130 (±0.057)	0.054 (±0.028)	3.983 (±1.002)
E3NN - Rotated Input	41 (±16)	86 (±11)	95 (±5)	0.129 (±0.058)	0.054 (±0.028)	3.987 (±1.020)
VoxMol - Unrotated Input	72 (±29)	81 (±31)	57 (±20)	0.188 (±0.071)	0.058 (±0.030)	3.567 (±1.348)
VoxMol - Rotated Input	71 (±29)	81 (±31)	57 (±20)	0.187 (±0.071)	0.058 (±0.030)	**3.528** (±1.250)
VoxMol (No Aug.) - Unrotated Input	**66** (±25)	84 (±22)	75 (±19)	0.187 (±0.069)	0.057 (±0.030)	3.600 (±1.449)
VoxMol (No Aug.) - Rotated Input	61 (±23)	85 (±20)	**83** (±17)	**0.176** (±0.070)	**0.056** (±0.029)	**3.379** (±1.272)
VoxMol (28 M) - Unrotated Input	**67** (±22)	84 (±21)	77 (±17)	0.169 (±0.064)	0.057 (±0.030)	3.244 (±1.199)
VoxMol (28 M) - Rotated Input	64 (±21)	84 (±21)	**81** (±15)	**0.163** (±0.064)	**0.055** (±0.029)	**3.154** (±1.204)
VoxMol (7 M) - Unrotated Input	**53** (±18)	84 (±15)	90 (±11)	0.165 (±0.060)	0.059 (±0.030)	3.469 (±1.029)
VoxMol (7 M) - Rotated Input	49 (±18)	84 (±15)	**92** (±10)	0.159 (±0.059)	0.057 (±0.029)	3.424 (±0.989)
VoxMol 50 % Train Subset - Unrotated Input	**75** (±30)	83 (±30)	46 (±20)	0.196 (±0.074)	0.058 (±0.031)	3.944 (±1.677)
VoxMol 50 % Train Subset - Rotated Input	70 (±29)	84 (±28)	**53** (±19)	**0.186** (±0.075)	**0.056** (±0.030)	**3.531** (±1.400)

Table 1. Quantitative metrics (Vignac et al., 2023) are very similar for generations regardless whether the seed molecule is rotated or not with E3NN and VoxMol. But VoxMol models trained with no augmentations ("VoxMol (No Aug.)"), smaller models ("VoxMol (28M)" and "VoxMol (7M)") or trained with less data ("VoxMol (50% Train Subset)") exhibit lower molecular stability and higher uniqueness when inputs are rotated, suggesting they are not equivariant during generation and that their generations are lower quality.

molecules per experiment. We compare generations across four models: E3NN (equivariant by design), full-size VoxMol, VoxMol trained with no augmentations, a smaller VoxMol (with 7 M and 28M parameters), and VoxMol trained on only 50% of the dataset.

Generation Quality and Diversity. To measure robustness to rotations, we compare commonly used generation quality metrics, such as stability, validity, uniqueness, and atom and bond distributions reported in Table 1. Generations with E3NN and full-size VoxMol are largely unaffected by seed molecule rotations, yielding similar stability and uniqueness metrics across experiments with and without rotating the

Model	TPSA ↓	LogP ↓	Fraction CSP3 ↓	Heavy Atoms ↓	Aromatic Rings ↓	H-Bond Donors ↓
E3NN	**0.0010**	**0.0013**	**0.0024**	**0.0023**	**0.0010**	**0.0001**
VoxMol (111 M)	**0.0025**	**0.0018**	**0.0029**	**0.0035**	**0.0007**	**0.0005**
VoxMol (No Aug.)	0.0222	0.0243	0.0203	0.0557	0.0104	0.0038
VoxMol (28 M)	0.0288	0.0118	0.0177	0.0308	0.0023	0.0037
VoxMol (50 %)	0.0495	0.0492	0.0701	0.0729	0.0039	0.0127

Table 2. KL Divergence between chemical property distributions of generated molecules with and without rotating the seed molecule. E3NN and VoxMol have the lowest KL divergence, showing that the generated distributions are very similar and not affected by rotations. On the other hand, VoxMol trained with no augmentations ("VoxMol (No Aug.)"), smaller model ("VoxMol (28M)") and with less training data ("VoxMol (50% Train Subset)") generate more different distributions with higher KL divergence between generated properties.

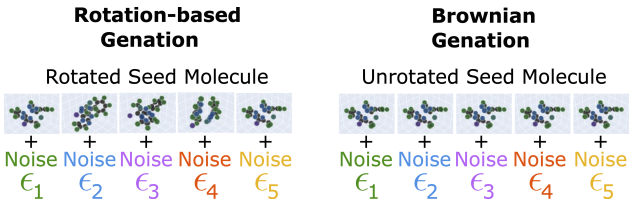

Figure 7. Experimental set up of generation equivariance. We compare generated molecules when the seed lead molecule is rotated to when the seed molecule is not rotated but noised with the same noise.

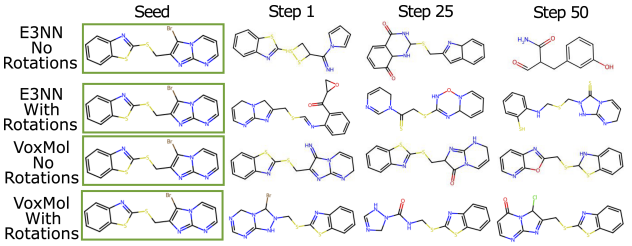

Figure 8. Generated molecules with and without rotating the seed molecule obtained with E3NN and VoxMol trained with rotations.

seed molecule. In contrast, models trained with no augmentations ("VoxMol (No Aug.)"), smaller model capacity ("VoxMol (28M)" and "VoxMol (7M)"), or reduced training data ("VoxMol (50% Train Subset)") exhibit decreased stability and increased uniqueness under rotation. These results suggest that while VoxMol can learn equivariance during reconstruction, it may fail to generalize this robustness to generation without sufficient capacity and data augmentation. Figure 8 shows example generated molecules.

Property Distribution Shifts Under Rotation. We also computed chemical properties of generated molecules using RDKit (Landrum et al., 2016) and computed the Kullback–Leibler (KL) divergence between the distributions of each property generated with and without rotations of the seed molecule. As seen in Table 2 E3NN and VoxMol show low divergence between rotated and unrotated generations, indicating their generations are very similar across orientations. However, models trained with no augmentation, smaller model size, and limited data all show higher KL divergence, revealing that rotations meaningfully alter their generation behavior. This suggests that these models are not fully equivariant during generation. We visualize histograms of properties of generated molecules with and without rotating the seed molecule obtained with VoxMol in Figure 9.

4.3. Property Predictions

To evaluate how input rotations affect property prediction, we train each model to predict asphericity, a geometry-sensitive property that measures how close a 3D object is to a perfect sphere. We add a predictive head to each model and compare the full-size VoxMol model (architecture details are provided in the Appendix), trained without rotational augmentations and a smaller model. For evaluation, we compute Spearman correlation and Mean Absolute Error (MAE) between predicted and ground truth values (Spearman-GT, MAE-GT). To test rotation robustness, we compute the same metrics between predictions on rotated versions of the same molecule (Spearman-Rot, MAE-Rot). If a model is equivariant, predictions should remain consistent across rotations (Gerken & Kessel, 2024).

We compare training with encoders only, where each model is trained to predict just the property ("Encoder-only"), and where each model is jointly trained to reconstruct the input molecule with an additional MSE voxel denoising loss used in addition to predicting the property ("Enc+Dec+Denoise"). All inputs are noised with the same noise as during training for reconstruction and during generation to better analyze the effect of the additional reconstruction (denoising) loss. Our results show that adding a reconstruction loss consistently improves both prediction accuracy and rotation robustness, suggesting it encourages the model to encode more spatially meaningful features (see Table 3). VoxMol trained without augmentation performs the worst and is clearly non-equivariant. Interestingly, a smaller VoxMol 7M model performs nearly as well as the full 111M parameter model, highlighting that model size is not the limiting factor for learning equivariant predictions unlike during generation.

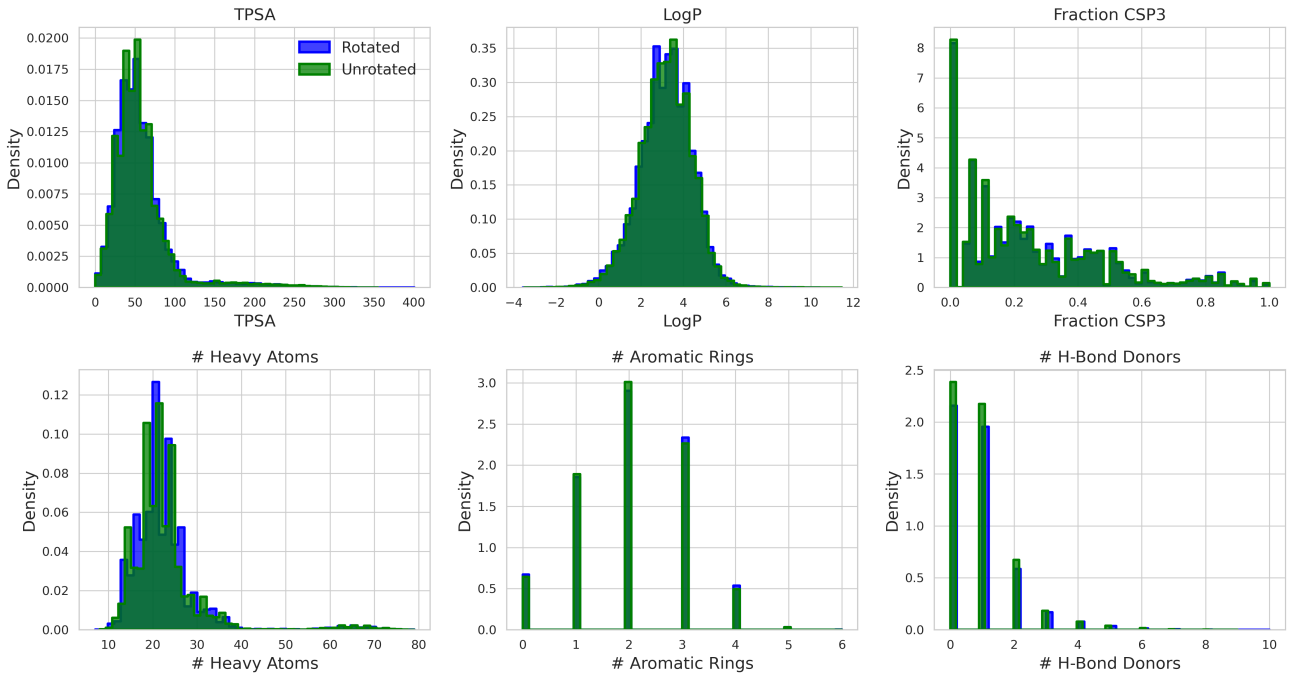

Figure 9. Distribution of chemical properties of generated molecules with and without rotating the seed molecule obtained with VoxMol trained with rotations.

Model	Setup	Spearman (GT) ↑	Spearman (Rot) ↑	MAE (GT) ↓	MAE (Rot) ↓
VoxMol (111 M)	Encoder-only	0.83	0.95	0.054	0.021
VoxMol (111 M)	Enc+Dec+Denoise	0.84	0.96	0.046	0.019
VoxMol (No Aug.)	Encoder-only	0.77	0.43	0.068	0.150
VoxMol (No Aug.)	Enc+Dec+Denoise	0.89	0.64	0.035	0.130
VoxMol (7 M)	Encoder-only	0.84	0.93	0.050	0.028
VoxMol (7 M)	Enc+Dec+Denoise	0.84	0.94	0.047	0.025

Table 3. Property (asphericity) prediction equivariance under input rotations. Spearman correlation and Mean Absolute Error (MAE) are computed between predicted and ground truth values (Spearman-GT, MAE-GT) and between predictions on rotated versions of the same molecule (Spearman-Rot, MAE-Rot).

5. Conclusions

Our results show that CNNs trained with rotation augmentation easily learn equivariance for reconstruction of noised molecules, even with small models, limited data and few training epochs. However, achieving robustness to rotations during generation requires larger models, larger datasets, and longer training.

Since generative models are trained only with a simple reconstruction loss which saturates easily, it may not fully allow the model to learn features useful for good performance on generation and prediction tasks. Therefore, large models and training datasets may be required because the model may need to learn redundant features. This is illustrated by very different latent embeddings of the same molecule at different rotations.

The training process could be improved with additional auxiliary tasks and regularization that would encourage the model to learn similar latent embeddings for the same molecules under different rotations. This may not only improve the model's equivariance during generation and prediction but perhaps could also improve the quality of generations by adding inductive bias and reduce the need for very large models that require long training times and large datasets due to their large number of parameters.

6. Impact Statement

This work aims to advance Machine Learning by studying learned equivariance in molecular generative models. Our findings may simplify model design and improve scalability for applications in drug discovery and materials science. We do not anticipate any specific ethical or societal risks.

References

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

Bao, F., Zhao, M., Hao, Z., Li, P., Li, C., and Zhu, J. Equivariant energy-guided sde for inverse molecular design. *arXiv preprint arXiv:2209.15408*, 2022.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on learning theory*, pp. 300–323. PMLR, 2018.

Cui, X., Mittal, A., Lu, S., Zhang, W., Saon, G., and Kingsbury, B. Soft random sampling: A theoretical and empirical analysis. *arXiv preprint arXiv:2311.12727*, 2023.

Diaz, I., Geiger, M., and McKinley, R. I. An end-to-end se (3)-equivariant segmentation network. *arXiv preprint arXiv:2303.00351*, 2023.

Fei, J. and Deng, Z. Rotation invariance and equivariance in 3d deep learning: a survey. *Artificial Intelligence Review*, 57(7):168, 2024.

Gebauer, N., Gastegger, M., and Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

Gerken, J., Carlsson, O., Linander, H., Ohlsson, F., Petersson, C., and Persson, D. Equivariance versus augmentation for spherical images. In *International Conference on Machine Learning*, pp. 7404–7421. PMLR, 2022.

Gerken, J. E. and Kessel, P. Emergent equivariance in deep ensembles. *arXiv preprint arXiv:2403.03103*, 2024.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887, 2022.

Huang, H., Sun, L., Du, B., and Lv, W. Learning joint 2-d and 3-d graph diffusion models for complete molecule generation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Kaufman, B., Williams, E. C., Pederson, R., Underkoffler, C., Panjwani, Z., Wang-Henderson, M., Mardirossian, N., Katcher, M. H., Strater, Z., Grandjean, J.-M., et al. Latent diffusion for conditional generation of molecules. *bioRxiv*, pp. 2024–08, 2024.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

Kvinge, H., Emerson, T., Jorgenson, G., Vasquez, S., Doster, T., and Lew, J. In what ways are deep neural networks invariant and how should we measure this? *Advances in Neural Information Processing Systems*, 35:32816–32829, 2022.

Landrum, G. et al. Rdkit: Open-source cheminformatics software. 2016.

Nowara, E., Pinheiro, P. O., Mahajan, S. P., Mahmood, O., Watkins, A. M., Saremi, S., and Maser, M. Nebula: Neural empirical bayes under latent representations for efficient and controllable design of molecular libraries. In *ICML 2024 AI for Science Workshop*, 2024.

Pinheiro, P. O., Rackers, J., Kleinhenz, J., Maser, M., Mahmood, O., Watkins, A. M., Ra, S., Sresht, V., and Saremi, S. 3d molecule generation by denoising voxel grids. *Advances in Neural Information Processing Systems*, 2023.

Pinheiro, P. O., Jamasb, A., Mahmood, O., Sresht, V., and Saremi, S. Structure-based drug design by denoising voxel grids. *arXiv preprint arXiv:2405.03961*, 2024.

Quiroga, F., Ronchetti, F., Lanzarini, L., and Bariviera, A. F. Revisiting data augmentation for rotational invariance in convolutional neural networks. In *Modelling and Simulation in Management Sciences: Proceedings of the International Conference on Modelling and Simulation in Management Sciences (MS-18)*, pp. 127–141. Springer, 2020.

Ragoza, M., Masuda, T., and Koes, D. R. Learning a continuous representation of 3d molecular structures with deep generative models. *arXiv preprint arXiv:2010.08687*, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Saremi, S. and Hyvärinen, A. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019.

Skalic, M., Jiménez, J., Sabbadin, D., and De Fabritiis, G. Shape-based generative modeling for de novo drug design. *Journal of chemical information and modeling*, 59(3):1205–1214, 2019.

Vignac, C., Osman, N., Toni, L., and Frossard, P. Midi: Mixed graph and 3d denoising diffusion for molecule generation. *arXiv preprint arXiv:2302.09048*, 2023.

Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. S. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 31, 2018.

Xu, M., Powers, A. S., Dror, R. O., Ermon, S., and Leskovec, J. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

A. Appendix

A.1. Model Architecture

We train two models to evaluate learned equivariance.

VoxMol is based on a U-Net model with 3D CNN layers with 4 levels of resolution and self-attention on the lowest two resolutions (Pinheiro et al., 2023). VoxMol was trained until convergence for 239 epochs with noise level of $\sigma = 0.9$ and learning rate of 1e-5. We trained VoxMol with data augmentation by randomly rotating and translating every sample in each iteration. See Pinheiro et al. (2023) for architecture and training details.

E3NN model is an equivariant version of VoxMol with E3NN layers. We use the SE(3)-equivariant 3D U-Net using steerable CNNs (Weiler et al., 2018) and we use the official implementation of Diaz et al. (2023). We tune the network hyperparameters for the best denoising performance, however, the E3NN model is not able to match the performance of VoxMol on denoising or generation metrics. We trained E3NN until convergence for 138 epochs and learning rate of 1e-2.

Predictive Head We add a lightweight Multi-Layer Perceptron (MLP) head to the UNet encoder for property prediction. The MLP includes average pooling to reduce spatial dimensions, followed by three linear layers with layer normalization, dropout of 0.1, and Shifted SoftPlus activations. We use Tanhshrink as the activation in the final layer for the regression task. The input voxels were noised with $\sigma = 0.9$ as in the reconstruction task.

A.2. Dataset

We train all models on a standard public dataset used for molecule generation called GEOM-drugs (GEOM) (Axelrod & Gomez-Bombarelli, 2022), following the train, validation, and test set splits used in (Vignac et al., 2023) with 1.1M/146K/146K molecules each. We train every model with soft random subsampling (Cui et al., 2023) and train on 10% of the training set in each epoch. We use voxel grids of dimension 64, and 8 atom channels ([C, H, O, N, F, S, Cl, Br]) with atomic radii of 0.25 $\mathring{A}$ resolution.

A.3. Additional Reconstruction Results

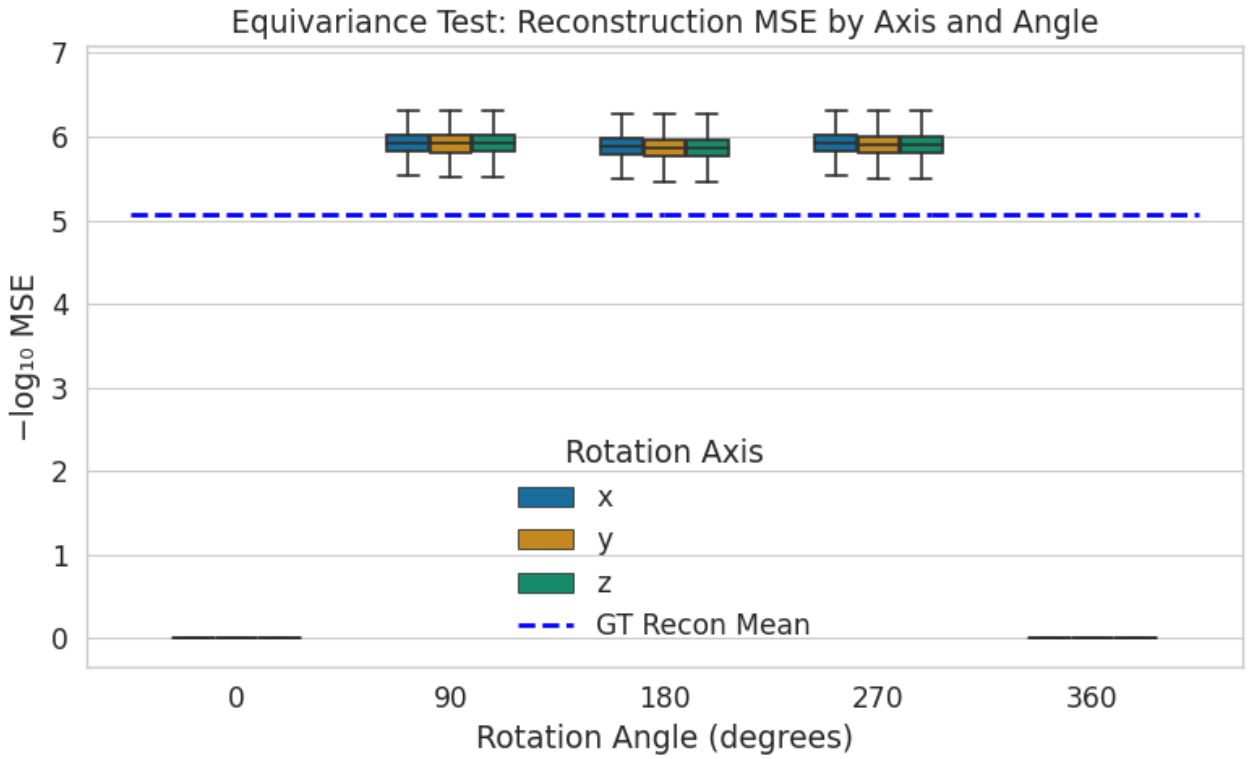

Figure 10. Rotational equivariance for the **111 M model**.

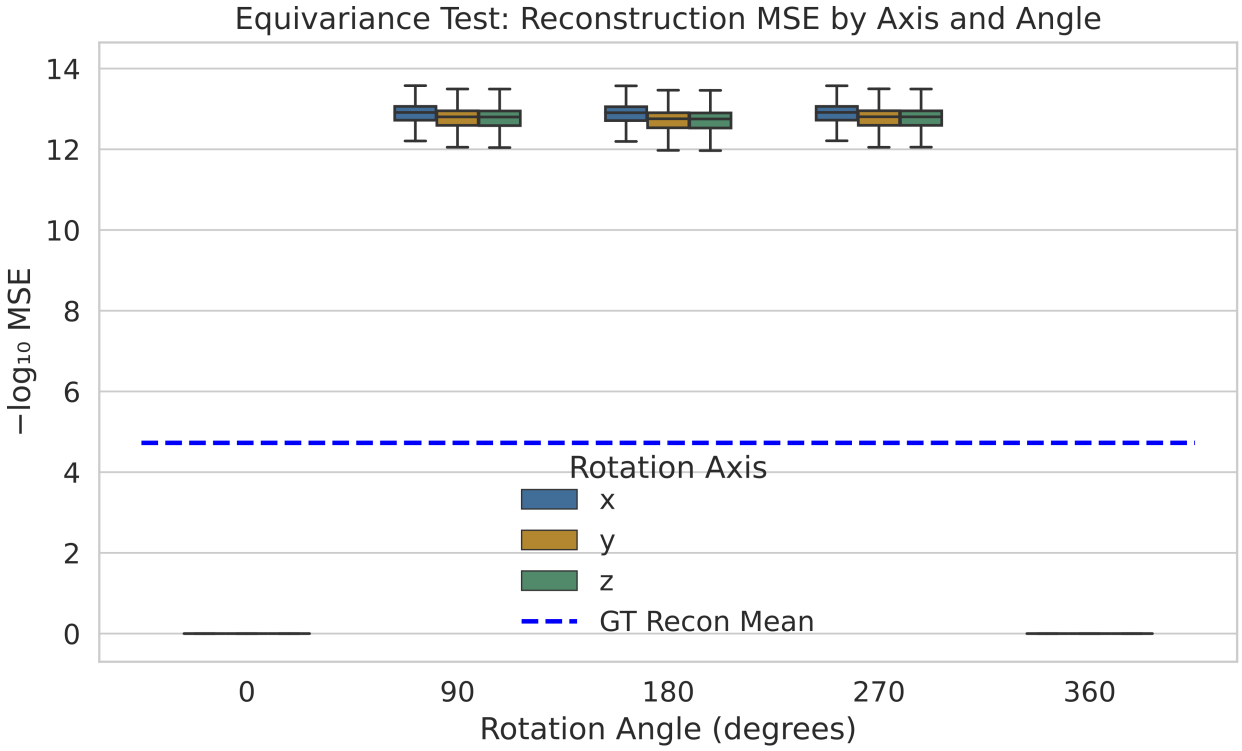

Figure 11. Rotational equivariance for the **E3NN model**.

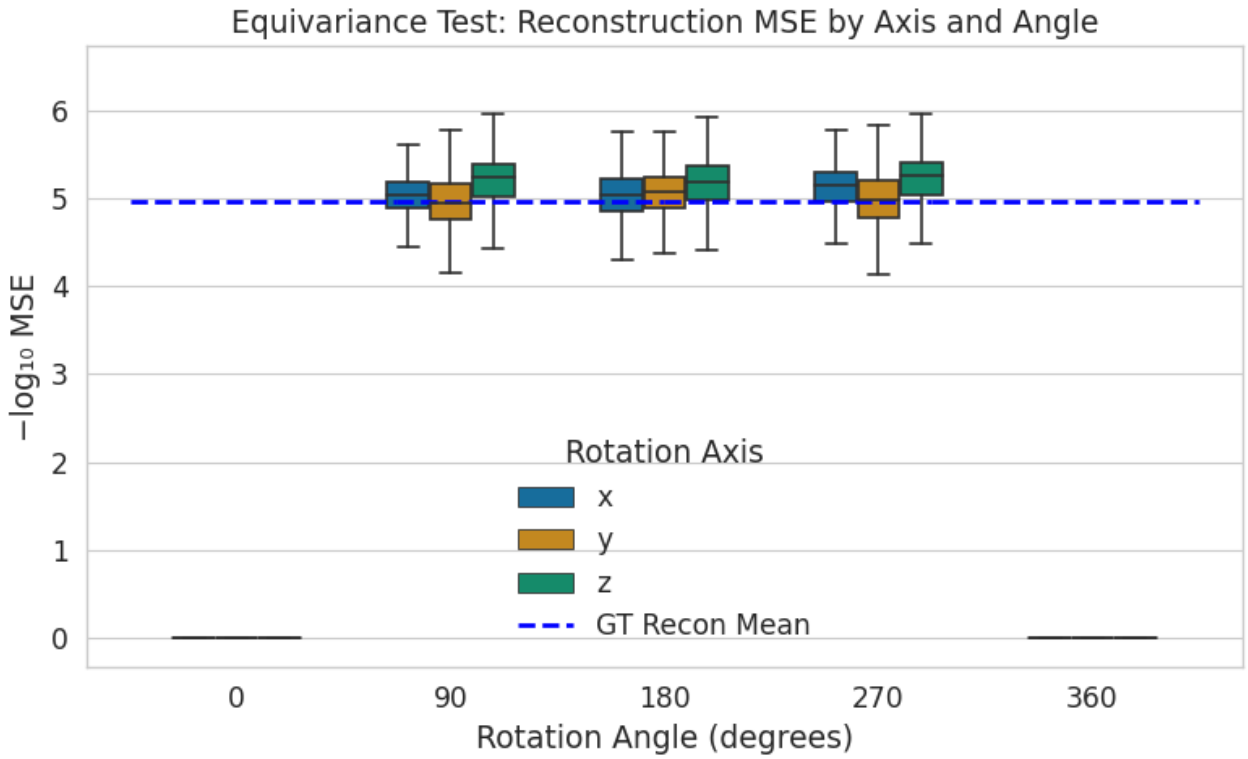

Figure 12. Rotational equivariance for the 111 M trained with **no rotation augmentations**.

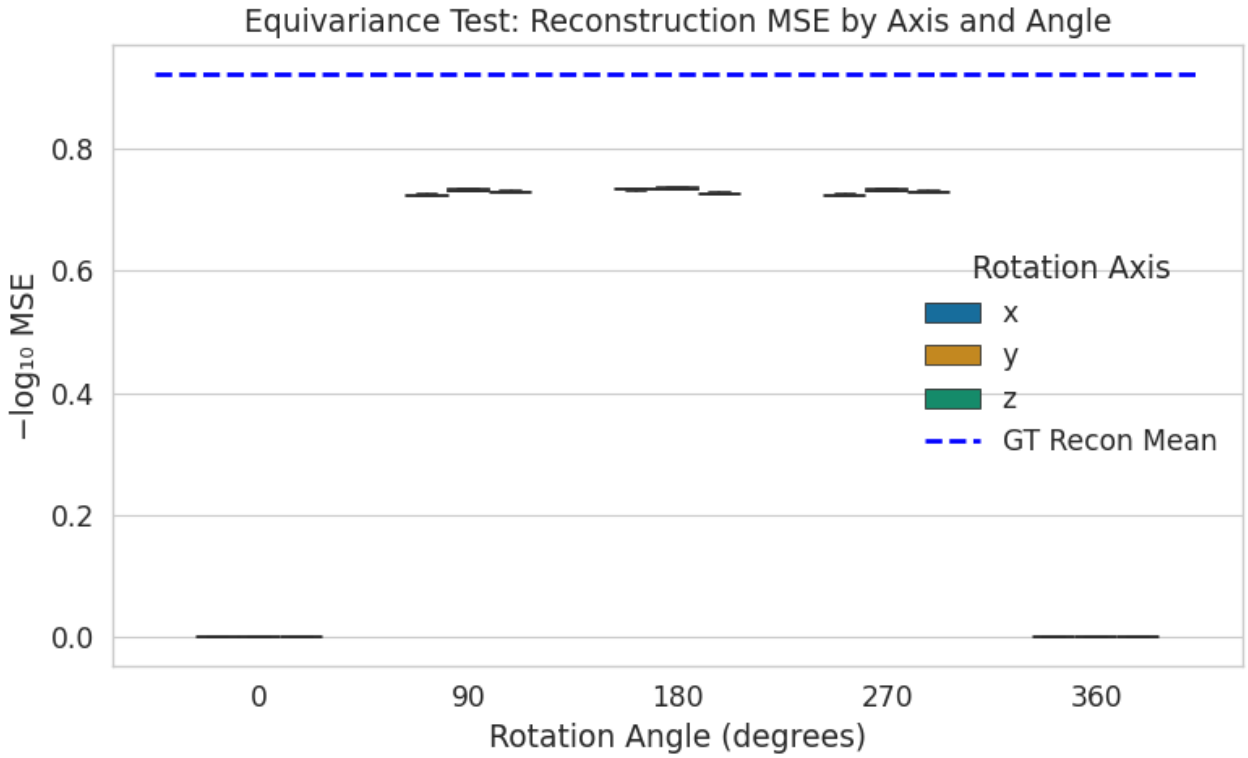

Figure 13. Rotational equivariance for the 111 M model with **randomly initialized weights** before training.

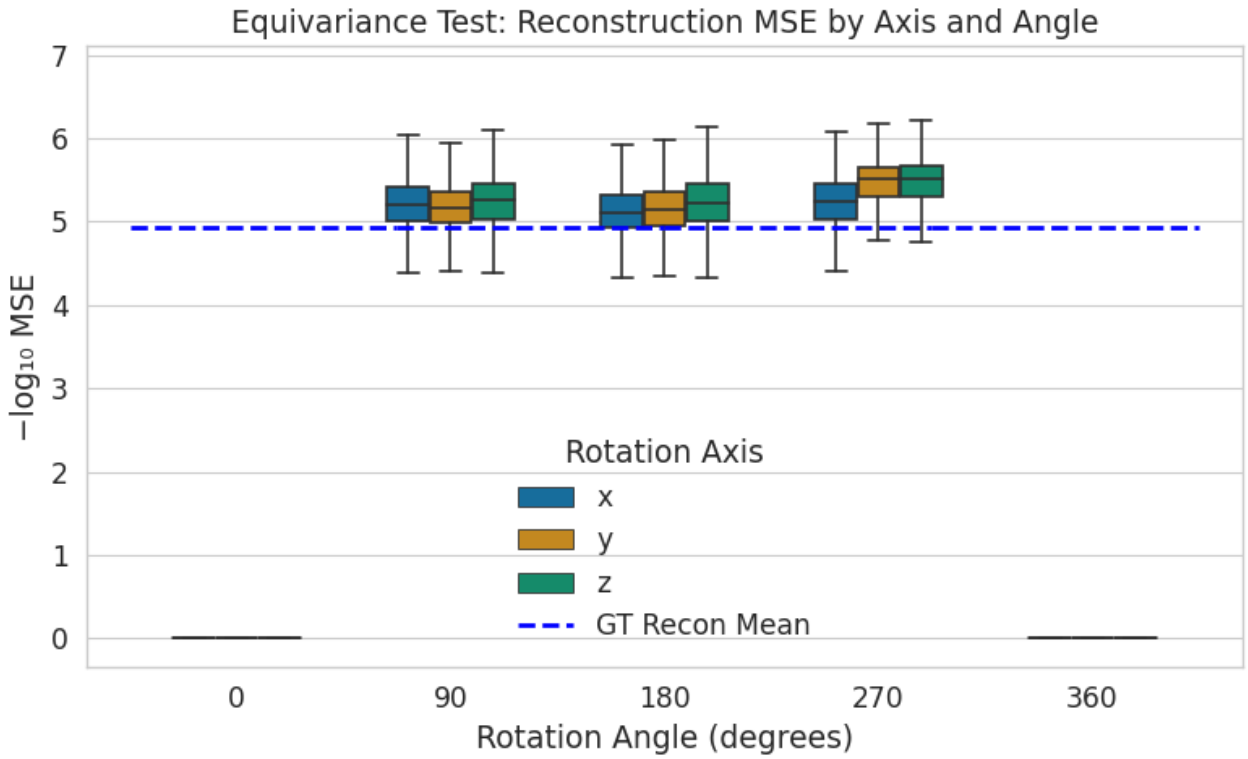

Figure 14. Rotational equivariance for the **28 M model**.

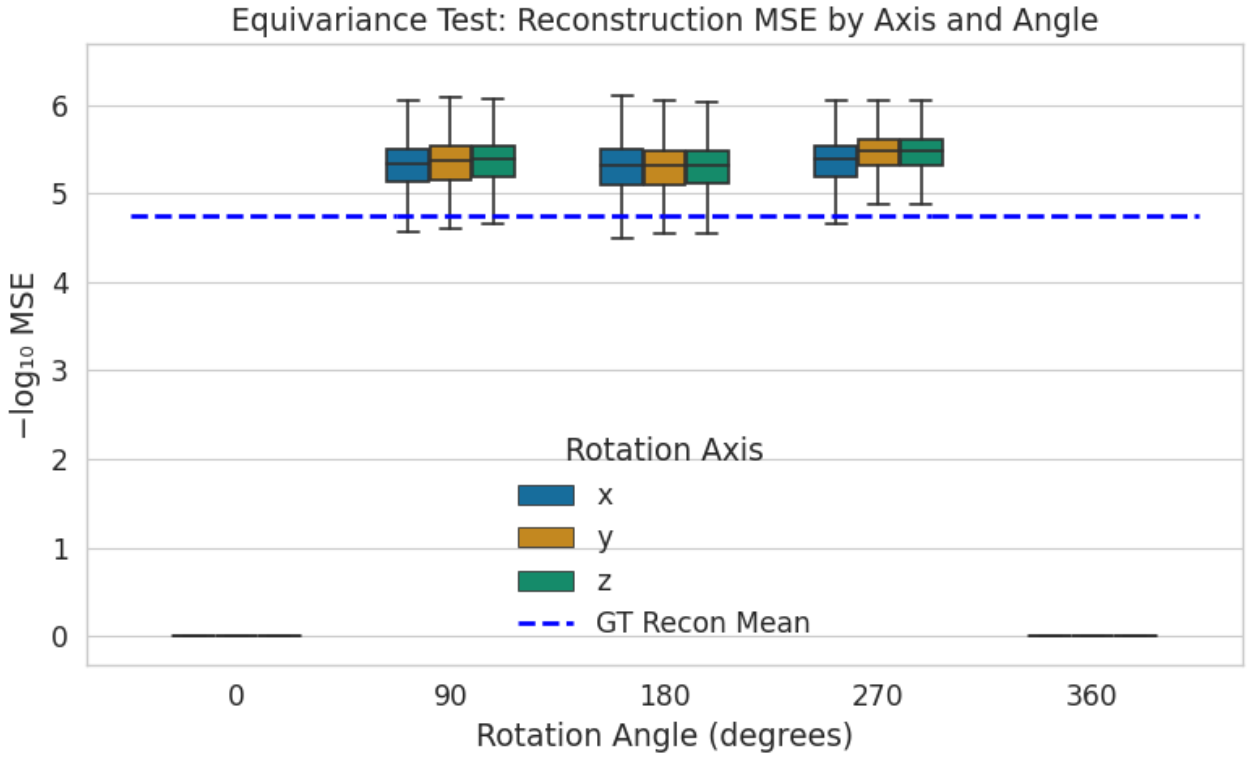

Figure 15. Rotational equivariance for the **7 M model**.

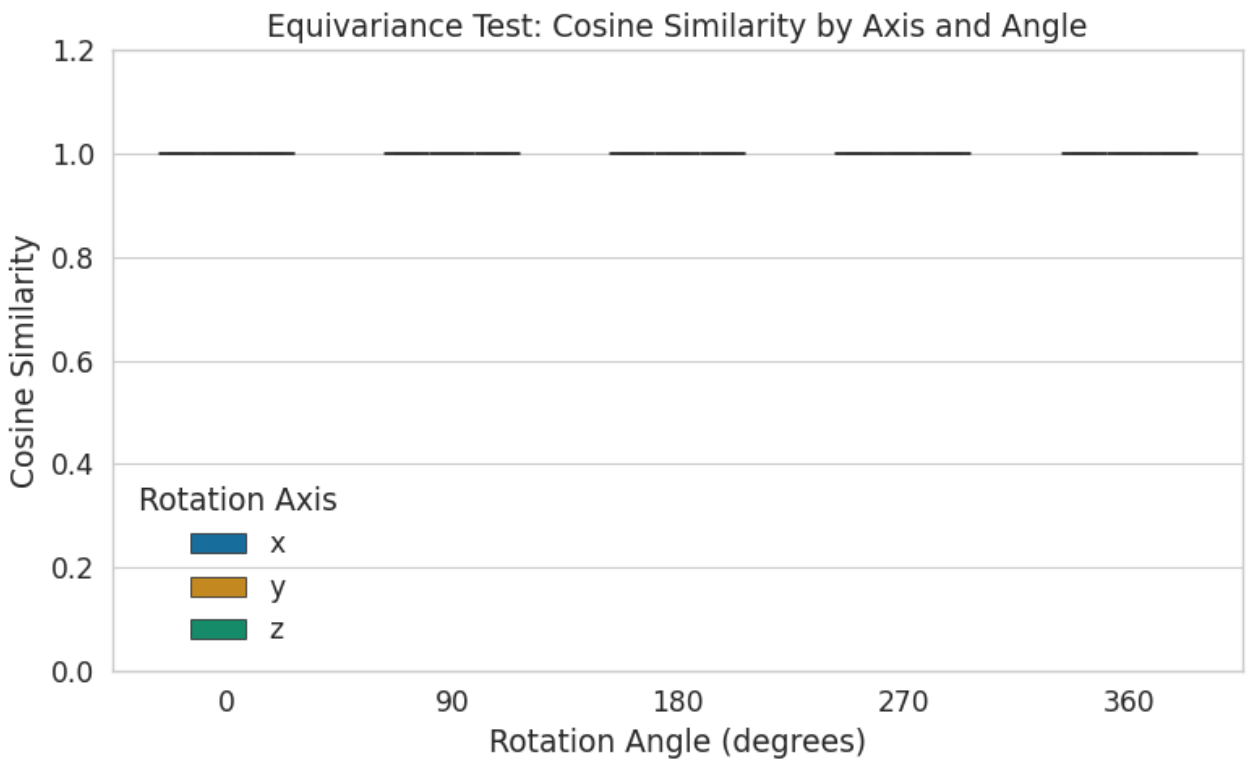

Figure 16. Cosine similarity between latent embeddings of the same molecules under different rotations for the **E3NN model**.

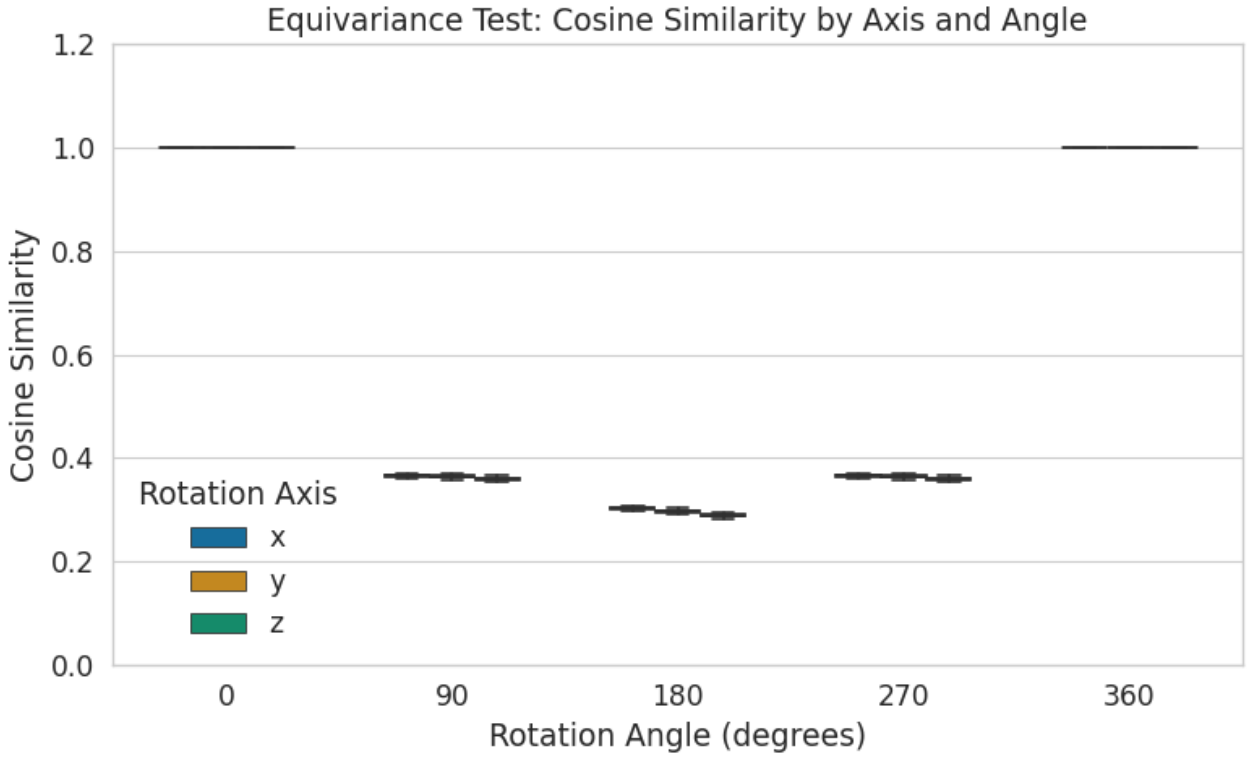

Figure 17. Cosine similarity between latent embeddings of the same molecules under different rotations for the 111 M model with **randomly initialized weights** before training.

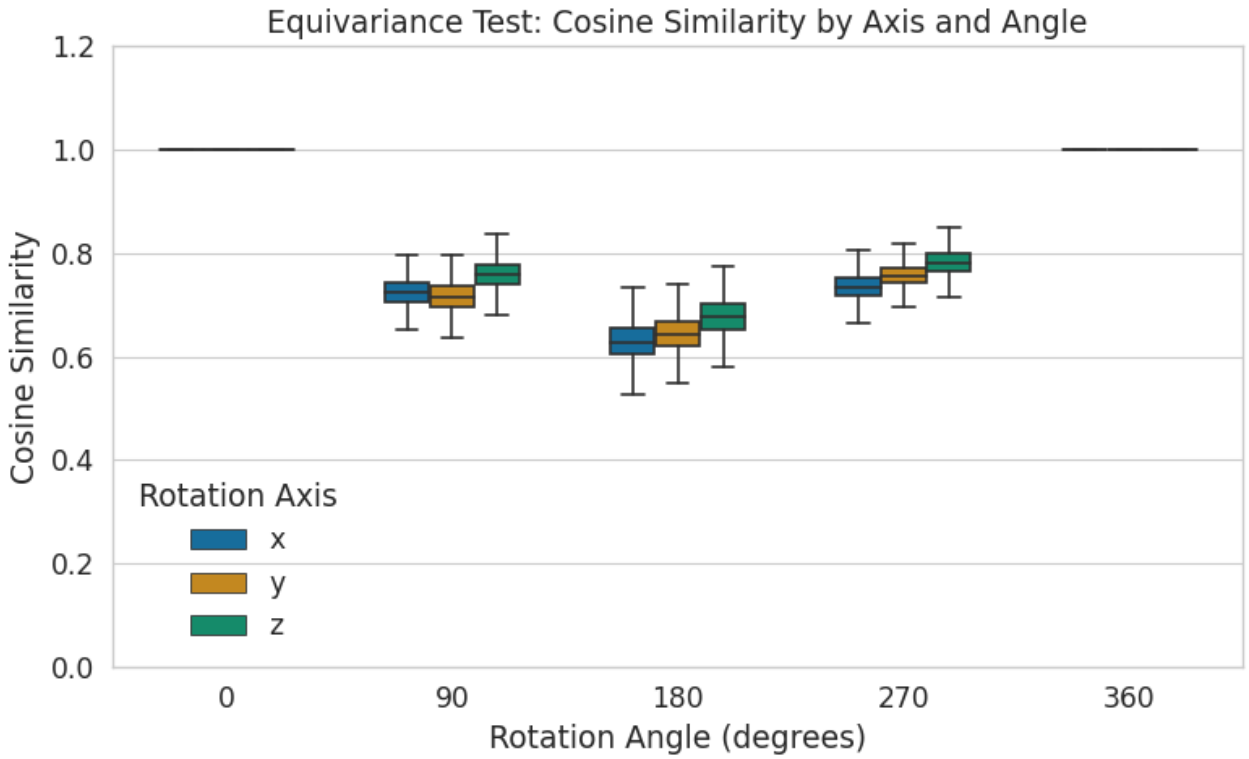

Figure 18. Cosine similarity between latent embeddings of the same molecules under different rotations for the **28 M model**.

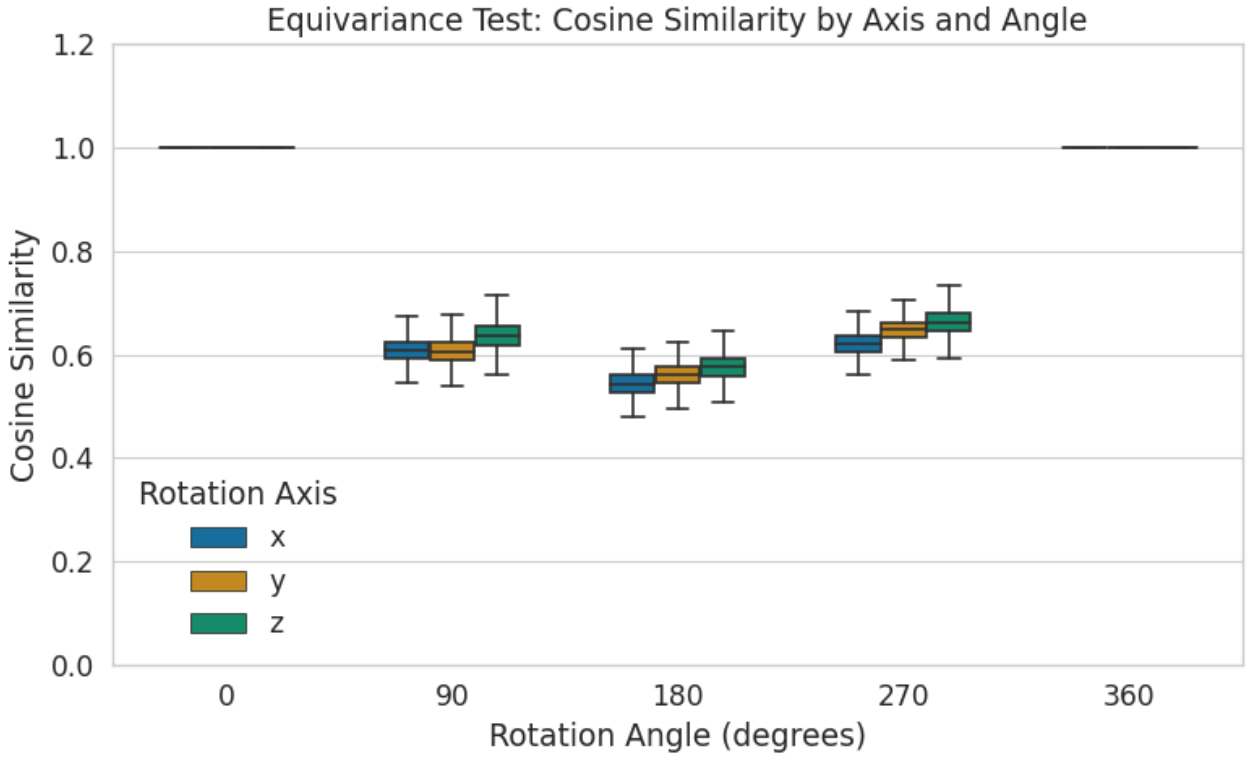

Figure 19. Cosine similarity between latent embeddings of the same molecules under different rotations for the **7 M model**.

	stable mol. ↑	stable atom ↑	valid %↑	unique %↑	valency W1↓	atom TV↓	bond TV↓	bond len W1↓	bond ang W1↓
E3NN - Unrotated Input	41 (±16)	97 (±1)	85 (±11)	95 (±6)	28 (±2)	0.130 (±0.057)	0.054 (±0.028)	0.004 (±0.001)	3.983 (±1.002)
E3NN - Rotated Input	41 (±16)	97 (±1)	86 (±11)	95 (±5)	28 (±2)	0.129 (±0.058)	0.054 (±0.028)	0.004 (±0.001)	3.987 (±1.020)
VoxMol - Unrotated Input	72 (±29)	99 (±1)	81 (±31)	57 (±20)	28 (±3)	0.188 (±0.071)	0.058 (±0.030)	0.005 (±0.001)	3.567 (±1.348)
VoxMol - Rotated Input	71 (±29)	99 (±1)	81 (±31)	57 (±20)	28 (±3)	0.187 (±0.071)	0.058 (±0.030)	0.005 (±0.001)	3.528 (±1.250)
VoxMol 111 M No Aug. - Unrotated Input	66 (±25)	99 (±1)	84 (±22)	75 (±19)	26 (±3)	0.187 (±0.069)	0.057 (±0.030)	0.004 (±0.001)	3.600 (±1.449)
VoxMol 111 M No Aug. - Rotated Input	61 (±23)	99 (±1)	85 (±20)	83 (±17)	25 (±3)	0.176 (±0.070)	0.056 (±0.029)	0.004 (±0.001)	3.379 (±1.272)

Table 4. Detailed seeded Generation Results with models initialized with lead molecules without rotations (Brownian Motion ("BM")) and with rotations ("rot.").

	stable mol. ↑	stable atom ↑	valid %↑	unique %↑	valency W1↓	atom TV↓	bond TV↓	bond len W1↓	bond ang W1↓
VoxMol - Unrotated Input	72 (±29)	99 (±1)	81 (±31)	57 (±20)	28 (±3)	0.188 (±0.071)	0.058 (±0.030)	0.005 (±0.001)	3.567 (±1.348)
VoxMol - Rotated Input	71 (±29)	99 (±1)	81 (±31)	57 (±20)	28 (±3)	0.187 (±0.071)	0.058 (±0.030)	0.005 (±0.001)	3.528 (±1.250)
VoxMol (28 M) - Unrotated Input	67 (±22)	99 (±1)	84 (±21)	77 (±17)	26 (±2)	0.169 (±0.064)	0.057 (±0.030)	0.004 (±0.001)	3.244 (±1.199)
VoxMol (28 M) - Rotated Input	64 (±21)	99 (±1)	84 (±21)	81 (±15)	26 (±2)	0.163 (±0.064)	0.055 (±0.029)	0.004 (±0.001)	3.154 (±1.204)
VoxMol (7 M) - Unrotated Input	53 (±18)	98 (±1)	84 (±15)	90 (±11)	29 (±2)	0.165 (±0.060)	0.059 (±0.030)	0.004 (±0.001)	3.469 (±1.029)
VoxMol (7 M) - Rotated Input	49 (±18)	98 (±1)	84 (±15)	92 (±10)	29 (±2)	0.159 (±0.059)	0.057 (±0.029)	0.004 (±0.001)	3.424 (±0.989)

Table 5. **Effect of model size** on seeded generation results with models initialized with lead molecules without rotations (Brownian Motion ("BM")) and with rotations ("rot."). Models were trained on 50 % training subset.

	stable mol. ↑	stable atom ↑	valid %↑	unique %↑	valency W1↓	atom TV↓	bond TV↓	bond len W1↓	bond ang W1↓
VoxMol 1 % - Unrotated Input	36 (±16)	97 (±1)	85 (±13)	98 (±5)	29 (±2)	0.171 (±0.053)	0.067 (±0.037)	0.004 (±0.001)	5.800 (±0.890)
VoxMol 1 % - Rotated Input	34 (±15)	96 (±1)	84 (±14)	98 (±5)	29 (±2)	0.169 (±0.052)	0.067 (±0.036)	0.004 (±0.001)	5.805 (±0.896)
VoxMol 10 % - Unrotated Input	73 (±27)	99 (±1)	84 (±27)	58 (±20)	27 (±3)	0.183 (±0.068)	0.058 (±0.030)	0.005 (±0.001)	3.745 (±1.576)
VoxMol 10 % - Rotated Input	69 (±26)	99 (±1)	84 (±26)	64 (±19)	27 (±2)	0.175 (±0.069)	0.056 (±0.029)	0.005 (±0.001)	3.487 (±1.420)
VoxMol 25 % - Unrotated Input	76 (±30)	99 (±1)	84 (±29)	44 (±19)	28 (±3)	0.195 (±0.075)	0.058 (±0.031)	0.005 (±0.001)	4.055 (±1.786)
VoxMol 25 % - Rotated Input	72 (±29)	99 (±1)	84 (±28)	51 (±20)	27 (±3)	0.185 (±0.075)	0.057 (±0.030)	0.005 (±0.001)	3.649 (±1.503)
VoxMol 50 % - Unrotated Input	75 (±30)	99 (±1)	83 (±30)	46 (±20)	28 (±4)	0.196 (±0.074)	0.058 (±0.031)	0.005 (±0.002)	3.944 (±1.677)
VoxMol 50 % - Rotated Input	70 (±29)	99 (±1)	84 (±28)	53 (±19)	27 (±3)	0.186 (±0.075)	0.056 (±0.030)	0.005 (±0.001)	3.531 (±1.400)

Table 6. **Effect of training set size** on seeded generation results with models initialized with lead molecules without rotations (Brownian Motion ("BM")) and with rotations ("rot."). Models were trained on 50 % training subset.

A.4. Additional Generation Results

	stable mol. ↑	stable atom ↑	valid %↑	unique %↑	valency W1↓	atom TV↓	bond TV↓	bond len W1↓	bond ang W1↓
VoxMol 50 epochs - Unrotated Input	63 (±20)	99 (±1)	85 (±18)	84 (±14)	25 (±2)	0.158 (±0.056)	0.057 (±0.030)	0.004 (±0.001)	3.250 (±1.209)
VoxMol 50 epochs - Rotated Input	60 (±20)	99 (±1)	85 (±18)	87 (±12)	25 (±2)	0.153 (±0.056)	0.055 (±0.028)	0.004 (±0.001)	3.193 (±1.281)
VoxMol 100 epochs - Unrotated Input	69 (±23)	99 (±1)	84 (±23)	73 (±18)	26 (±3)	0.169 (±0.062)	0.057 (±0.030)	0.004 (±0.001)	3.330 (±1.348)
VoxMol 100 epochs - Rotated Input	66 (±23)	99 (±1)	84 (±22)	76 (±16)	26 (±2)	0.164 (±0.064)	0.056 (±0.029)	0.004 (±0.001)	3.165 (±1.221)
VoxMol 200 epochs - Unrotated Input	70 (±28)	99 (±1)	84 (±28)	61 (±20)	27 (±3)	0.181 (±0.065)	0.057 (±0.029)	0.005 (±0.001)	3.459 (±1.376)
VoxMol 200 epochs - Rotated Input	66 (±26)	99 (±1)	83 (±26)	67 (±18)	27 (±2)	0.173 (±0.067)	0.055 (±0.028)	0.004 (±0.001)	3.290 (±1.343)
VoxMol 300 epochs - Unrotated Input	73 (±29)	99 (±1)	83 (±29)	54 (±21)	27 (±3)	0.187 (±0.070)	0.058 (±0.030)	0.005 (±0.001)	3.737 (±1.601)
VoxMol 300 epochs - Rotated Input	67 (±28)	99 (±1)	83 (±28)	61 (±20)	27 (±2)	0.178 (±0.072)	0.056 (±0.029)	0.005 (±0.001)	3.415 (±1.394)
VoxMol 400 epochs - Unrotated Input	75 (±28)	99 (±1)	85 (±28)	50 (±21)	28 (±4)	0.189 (±0.070)	0.058 (±0.030)	0.005 (±0.002)	3.935 (±1.680)
VoxMol 400 epochs - Rotated Input	70 (±28)	99 (±1)	85 (±26)	57 (±20)	27 (±3)	0.179 (±0.070)	0.056 (±0.030)	0.005 (±0.001)	3.496 (±1.381)
VoxMol 500 epochs - Unrotated Input	76 (±29)	99 (±1)	84 (±29)	47 (±20)	28 (±4)	0.194 (±0.073)	0.058 (±0.031)	0.005 (±0.002)	3.968 (±1.871)
VoxMol 500 epochs - Rotated Input	69 (±29)	99 (±1)	84 (±28)	53 (±19)	27 (±3)	0.184 (±0.074)	0.056 (±0.030)	0.005 (±0.001)	3.519 (±1.476)

Table 7. **Effect of number of training epochs** on seeded generation results with models initialized with lead molecules without rotations (Brownian Motion ("BM")) and with rotations ("rot."). Models were trained on 50 % training subset.