
Towards Unified and Lossless Latent Space for 3D Molecular Latent Diffusion Modeling

Yanchen Luo¹, Zhiyuan Liu^{2*}, Yi Zhao¹, Sihang Li¹, Hengxing Cai³,
Kenji Kawaguchi², Tat-Seng Chua², Yang Zhang², Xiang Wang^{1*}

¹University of Science and Technology of China

²National University of Singapore

³DP Technology

luoyanchen@mail.ustc.edu.cn

acharkq@gmail.com

xiangwang1223@gmail.com

Abstract

3D molecule generation is crucial for drug discovery and material science, requiring models to process complex multi-modalities, including atom types, chemical bonds, and 3D coordinates. A key challenge is integrating these modalities of different shapes while maintaining $SE(3)$ equivariance for 3D coordinates. To achieve this, existing approaches typically maintain separate latent spaces for invariant and equivariant modalities, reducing efficiency in both training and sampling. In this work, we propose Unified Variational Auto-Encoder for **3D** Molecular Latent Diffusion Modeling (**UAE-3D**), a multi-modal VAE that compresses 3D molecules into latent sequences from a unified latent space, while maintaining near-zero reconstruction error. This unified latent space eliminates the complexities of handling multi-modality and equivariance when performing latent diffusion modeling. We demonstrate this by employing the Diffusion Transformer—a general-purpose diffusion model without any molecular inductive bias—for latent generation. Extensive experiments on GEOM-Drugs and QM9 datasets demonstrate that our method significantly establishes new benchmarks in both *de novo* and conditional 3D molecule generation, achieving leading efficiency and quality. On GEOM-Drugs, it reduces FCD by 72.6% over the previous best result, while achieving over 70% relative average improvements in geometric fidelity. Our code is released at <https://github.com/lyc0930/UAE-3D/>.

1 Introduction

The discovery of novel molecules is fundamental to various scientific fields, particularly in drug and material development. Given significant progress has been made in designing 2D molecular graphs [1–4], recent research has increasingly focused on the generation of 3D molecules [5, 6]. Unlike 2D molecular generation, which focuses on forming valid molecular structures based on chemical bonds, 3D generation must also predict 3D atomic coordinates that align with the 2D structure. Accurate 3D molecule generation is essential to power many important applications, such as structure-based drug design [7] and inverse molecule design targeting quantum properties [8].

3D molecule generation is challenging due to its multi-modal nature. As shown in Figure 1(a), a 3D molecule consists of features of three distinct modalities: atom types, atomic coordinates, and chemical bonds. This requires the generation model to handle both discrete (*e.g.*, atom types) and continuous features (*e.g.*, coordinates), while also addressing differences in feature shapes (atom-wise

*Corresponding authors.

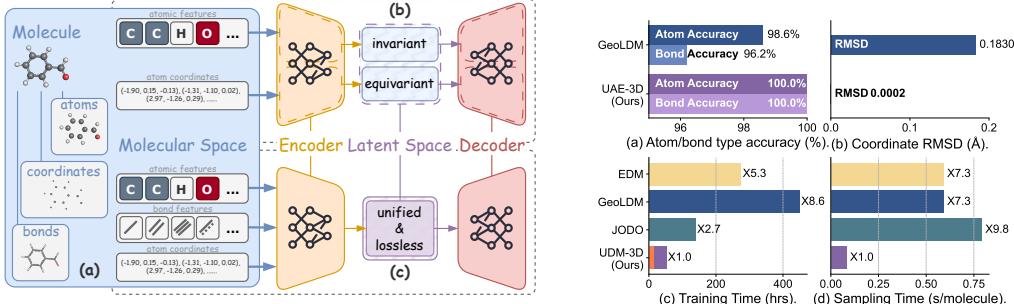


Figure 1: (a) A 3D molecule has multi-modal features. (b) Prior methods use separate latent spaces for equivariant (3D) and invariant (2D) modalities, inducing unnecessary complexity for model architecture. (c) UAE-3D reduces this complexity by establishing a unified and near-lossless latent space that integrates all molecular modalities.

v.s. edge-wise features). Worse still, the modality of 3D coordinate requires special care to respect rotational and translational equivariance, whereas other modalities do not, further complicating the task. Mitigating this multi-modal challenge, prior works mostly process each modality in separate latent spaces: some maintain separate diffusion processes for each modality [6, 9–11], while others interleave the prediction of different modalities across autoregressive generation steps [12, 13]. However, we argue that handling each modality separately complicates the model design, reducing both the training and sampling efficiency. Moreover, processing each modality separately risks compromising the consistency between them. These issues raise a critical question: *Can we design a unified generative model that seamlessly integrates all three modalities of 3D molecule generation?*

To answer this research question, we propose to build a multi-modal latent diffusion model (LDM) [14] for the unified generative modeling of 3D molecules. LDM extends the diffusion paradigm [15] by operating in a compressed latent space learned through variational autoencoders (VAEs) [16], offering improved computational efficiency and generation quality. For 3D molecules, we can build a multi-modal VAE that compresses all three modalities in a single unified latent space. Scrutinizing the previous works, we identify the following challenges to achieve this purpose:

- **The Challenge of Compressing Multi-modality in One Latent Space.** While previous studies have explored LDMs for 3D molecule and protein generation [17, 18], they fail to build a unified latent space that integrates all modalities. Their difficulty arises from the reliance on neural networks with baked-in 3D equivariance [19, 6], which mostly maintain separate latent spaces for 3D equivariant and invariant modalities (cf. Figure 1(b)).
- **The Challenge of Ensuring Near-Lossless Compression.** Ensuring a low reconstruction error for the molecular VAE is critical, as even small errors can result in invalid or unstable 3D structures. Furthermore, imprecise latents that cannot reproduce the original molecule can propagate their errors to the LDM, easily disrupting the generative modeling. Despite its critical importance, reconstruction error has been largely overlooked in previous works [17, 18] (cf. Figure 2(a;b)).

In this work, we introduce **UAE-3D**, Unified Variational Auto-Encoder for 3D Molecular Latent Diffusion Modeling, a VAE that can compress the multi-modal features of 3D molecules into a unified latent space while maintaining near zero (100% atom/bond accuracy and 2E-4 coordinate RMSD) reconstruction error. To obtain a latent space integrating both 3D equivariant and invariant modalities, we draw inspiration from the “bitter lesson” [20] of 3D molecules: rather than using models with intricate, baked-in 3D equivariance, UAE-3D trains a neural network to “learn” 3D equivariance through our tailor-made SE(3) augmentations, encouraging the transformation on the input coordinates to be reflected equivariantly on the output coordinates. Moreover, UAE-3D employs the Relational Transformer [21] as its encoder, leveraging its scalability and flexibility to incorporate both atom-wise and edge-wise features. A Transformer-based [22] decoder is jointly trained with the encoder to reconstruct both 3D invariant and equivariant molecular features.

Despite its simplicity, UAE-3D can compress 3D molecules into token sequences in a unified latent space, **eliminating the complexities of handling multi-modalities and 3D equivariance in latent diffusion modeling**. Figure 2(a;b) show that UAE-3D reaches near-lossless reconstruction of atom and bond types, with a near-zero coordinate RMSD. To further demonstrate its effectiveness, we

employ the Diffusion Transformer (DiT) [23], a general-purpose diffusion backbone without any molecular inductive bias, to model UAE-3D’s latents. We show that DiT can successfully generate stable and valid 3D molecules, with significantly improved training (by 2.7 times) and sampling (by 7.3 times) efficiency (*cf.* Figure 2(c;d)). We refer to this LDM pipeline as **UDM-3D: Unified Latent Diffusion Modeling for 3D Molecule Generation**.

Extensive experiments demonstrate that UDM-3D, powered by UAE-3D, achieves state-of-the-art results in both *de novo* and conditional 3D molecule generation on the QM9 [24] and GEOM-Drugs [25] datasets. On GEOM-Drugs, our model achieves a 72.6% reduction in Fréchet ChemNet Distance [3], while significantly reducing the MMD of bond length, bond angle, and dihedral angle by 88.4%, 55.6%, and 74.0%, respectively. In conditional generation on QM9, it achieves the lowest MAE in five out of six properties. These results confirm the effectiveness of our unified modeling approach. Ablation studies show the effectiveness of our key components. Further analysis reveals that UAE-3D’s latents present structured variations across geometric movement.

2 Background: 3D Molecular Latent Diffusion Models

A 3D molecular LDM involves a 3D molecular VAE to compress 3D molecules into the latent space, where a diffusion model performs generative modeling. Below, we introduce both components.

Notations. A 3D molecule is represented by $\mathbf{G} = \langle \mathbf{F}, \mathbf{E}, \mathbf{X} \rangle$, where $\mathbf{F} \in \mathbb{R}^{|\mathcal{V}| \times d_1}$ is the atom feature matrix (*e.g.*, atomic numbers), $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}| \times d_2}$ is the bond feature matrix (*e.g.*, bond connectivity and type), and $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times 3}$ is the 3D atom coordinate matrix. d_1 and d_2 are dimensions of atom and bond features. \mathcal{V} is \mathbf{G} ’s set of atoms and $|\mathcal{V}|$ denotes the number of atoms.

3D Molecular Variational Auto-Encoding. A molecular VAE consists of an encoder \mathcal{E} and a decoder \mathcal{D} . Given a 3D molecule \mathbf{G} , the encoder \mathcal{E} maps it into a sequence of latent tokens $\mathcal{E}(\mathbf{G}) = \mathbf{Z} = \{z_i \in \mathbb{R}^d | i \in \mathcal{V}\}$. Each latent z_i is sampled from a Gaussian distribution $\mathcal{N}(z_i; \mu_i, \sigma_i)$ using the reparameterization trick [16], where μ_i and σ_i are learned parameters for atom i . The decoder \mathcal{D} then reconstructs \mathbf{G} from these latents: $\hat{\mathbf{G}} = \mathcal{D}(\mathbf{Z})$. The VAE is trained with a reconstruction loss and a regularization term:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{\mathbf{G} \sim p_{\text{data}}} \left[\|\hat{\mathbf{G}} - \mathbf{G}\| + D_{\text{KL}}(q(\mathbf{Z}|\mathbf{G}) \| p(\mathbf{z})) \right], \quad (1)$$

where D_{KL} denotes the Kullback-Leibler divergence; $q(\mathbf{Z}|\mathbf{G}) = \prod_{i \in \mathcal{V}} \mathcal{N}(z_i; \mu_i, \sigma_i)$ is the approximated posterior distribution; and $\|\hat{\mathbf{G}} - \mathbf{G}\|$ measures the difference between the original and the predicted graph, whose definition varies across different works. In [17], $\|\mathbf{G} - \hat{\mathbf{G}}\|$ combines an MSE loss for 3D coordinates and a cross-entropy loss for atom types.

Diffusion Model. Building on the molecular VAE’s latent \mathbf{Z} , a diffusion model [14, 26] performs generative modeling. In forward diffusion process, we gradually add noise to the original latent $\mathbf{Z}^{(0)} = \mathbf{Z}$ following $\mathbf{Z}^{(t)} \sim \mathcal{N}(\mathbf{Z}^{(t)}; \sqrt{\bar{\alpha}^{(t)}} \mathbf{Z}^{(0)}, (1 - \bar{\alpha}^{(t)}) \mathbf{I})$, where $t \in (0, 1]$ is the diffusion timestep, and $\bar{\alpha}^{(t)}$ is a hyperparameter controlling the signal/noise ratio. In practice, $\mathbf{Z}^{(t)}$ is sampled as $\mathbf{Z}^{(t)} = \sqrt{\bar{\alpha}^{(t)}} \mathbf{Z}^{(0)} + \sqrt{1 - \bar{\alpha}^{(t)}} \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Given the noised latents $\mathbf{Z}^{(t)}$, a diffusion model $\epsilon_{\theta}(\mathbf{Z}^{(t)}, t)$ is trained to predict the added noise ϵ by minimizing the MSE loss $\|\epsilon_{\theta}(\mathbf{Z}^{(t)}, t) - \epsilon\|^2$. Once trained, new latents $\hat{\mathbf{Z}}^{(0)}$ are sampled with the model ϵ_{θ} via iterative denoising [15], and the decoder \mathcal{D} reconstructs the corresponding 3D molecule via $\hat{\mathbf{G}} = \mathcal{D}(\hat{\mathbf{Z}}^{(0)})$.

Separated Latent Spaces. The VAE latents in previous works [18, 17] include two parts: $\mathbf{Z} = [\bar{\mathbf{Z}}; \tilde{\mathbf{Z}}]$, where $\tilde{\mathbf{Z}}$ is for equivariant features and $\bar{\mathbf{Z}}$ is for 3D invariant features. They also rely on diffusion models with baked-in 3D equivariance. In contrast, as Section 3 shows, our method uses a unified latent space, allowing a general-purpose diffusion model—without any geometric or molecular inductive bias—to achieve strong performance.

3 Methodology

In this section, we propose the **UAE-3D**, a multi-modal variational auto-encoder designed to effectively compress the diverse modalities of 3D molecules into a unified latent space. Based on UAE-3D, we introduce **UDM-3D**, an LDM for 3D molecule generation.

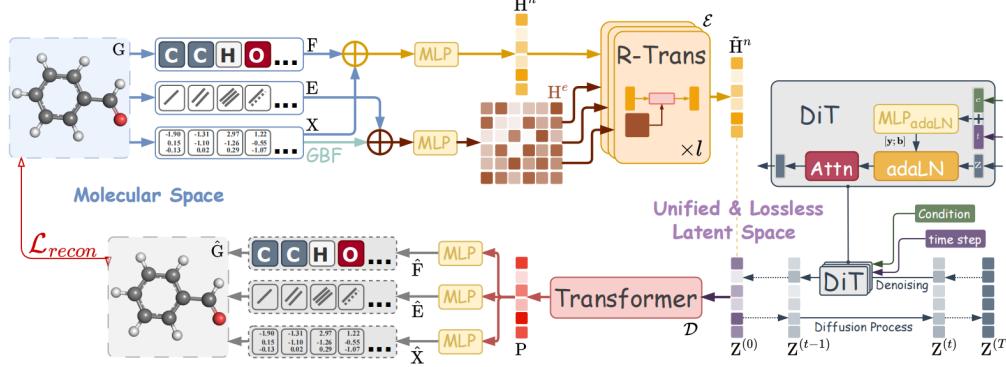


Figure 3: Overview of the UDM-3D and UAE-3D models. The UAE-3D encodes 3D molecules from molecular space into a unified latent space, integrating multi-modal features such as atom types, chemical bonds, and 3D coordinates. Utilizing this latent space, UDM-3D employs a DiT to perform generative modeling. Then, the denoised latents are decoded back into 3D molecules.

3.1 Unified Variational Auto-Encoder for 3D Molecular Latent Diffusion Modeling

UAE-3D is designed to address the complexities of multi-modal and equivariance of molecular data by compressing the atomic features, bond features, and atomic coordinates into a unified latent space. This is achieved through three key components: (1) a Relational-Transformer [21] that effectively integrates the multi-modal features into token sequences, (2) a Transformer-based [22] decoder and the reconstruction loss for 3D molecule reconstruction; and (3) our tailor-made SE(3)-equivariant data augmentations to train the model to learn 3D equivariance.

Compressing 3D Molecules with Relational Transformer (R-Trans) Encoder. Given a 3D molecular $\mathbf{G} = \langle \mathbf{F}, \mathbf{E}, \mathbf{X} \rangle$, we compute its initial atom-wise and edge-wise embeddings as:

$$\mathbf{H}^n = \text{MLP}([\mathbf{X}; \mathbf{F}]) \in \mathbb{R}^{|\mathcal{V}| \times d}, \quad (2) \quad \mathbf{H}^e = \text{MLP}([\mathbf{E}; \mathbf{D}]) \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}| \times d}, \quad (3)$$

where $\mathbf{D}_{ij} = \text{GBF}(\|\mathbf{X}_i - \mathbf{X}_j\|^2) \in \mathbb{R}^d$; $\text{GBF}(\cdot)$ implements the Gaussian basis functions to expand inter-atomic distances as feature vectors [27]; and d is the embedding dimension. The initial node embeddings \mathbf{H}^n combines both atomic features and its 3D coordinates, while the edge embeddings \mathbf{H}^e incorporate bond features and inter-atomic distances. \mathbf{H}^n and \mathbf{H}^e together effectively represent all 3D molecular information for the subsequent encoding. Given these embeddings, we process them with the molecular encoder $\mathcal{E}(\cdot)$ of L layers of R-Trans [21]:

$$\tilde{\mathbf{H}}^n = \text{R-Trans}(\mathbf{H}^n, \mathbf{H}^e) \in \mathbb{R}^{|\mathcal{V}| \times d}, \quad (4)$$

where $\tilde{\mathbf{H}}^n$ denotes the updated atom embedding for the next layer, computed as follows:

$$\mathbf{Q}_{ij} = [\mathbf{H}_i^n; \mathbf{H}_{ij}^e] \mathbf{W}^q \in \mathbb{R}^d, \quad (5) \quad [\mathbf{K}_{ij}; \mathbf{V}_{ij}] = [\mathbf{H}_j^n; \mathbf{H}_{ij}^e] \mathbf{W}^{kv} \in \mathbb{R}^{2d}, \quad (6)$$

$$\alpha_{ij} = \text{softmax}_j \left(\frac{\mathbf{Q}_{ij} \mathbf{K}_{ij}^\top}{\sqrt{d}} \right) \in \mathbb{R}, \quad (7) \quad \tilde{\mathbf{H}}_i^n = \text{MLP} \left(\sum_j \alpha_{ij} \mathbf{V}_{ij} \right) \in \mathbb{R}^d, \quad (8)$$

where \mathbf{W}^q and \mathbf{W}^{kv} are linear projectors. Unlike the original R-Trans, our implementation does not update the edge \mathbf{H}^e every layer, which improves efficiency without compromising performance.

Compared to Transformer [22], R-Trans is well-suited for 3D molecule encoding, that it can effectively integrate edge embeddings $\mathbf{H}^e \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}| \times d}$ and atom embeddings $\mathbf{H}^n \in \mathbb{R}^{|\mathcal{V}| \times d}$, despite their different shapes. This integration occurs during the computation of queries, keys and values (Equation (5) and (6)), ensuring the output atom embeddings fully incorporate edge information.

Decoder. Given encoder’s latents \mathbf{Z} , we employ a Transformer [22] decoder $\mathcal{D}(\cdot)$ to obtain the atom-wise output $\mathbf{P} = \text{Transformer}(\mathbf{Z}) \in \mathbb{R}^{|\mathcal{V}| \times d}$. Unlike the encoder, our decoder includes no molecule-specialized design, because it processes sequences of latents. Further, we employ three MLP predictors for atom, bond types, and atom coordinates to reconstruct complete 3D molecules:

$$\hat{\mathbf{X}}_i = \text{MLP}_1(\mathbf{P}_i) \in \mathbb{R}^3; \quad \hat{\mathbf{F}}_i = \text{MLP}_2(\mathbf{P}_i) \in \mathbb{R}^{N_a}; \quad \hat{\mathbf{E}}_{ij} = \text{MLP}_3(\mathbf{P}_i + \mathbf{P}_j) \in \mathbb{R}^{N_b}, \quad (9)$$

where N_a and N_b are the number of atom and bond types.

Reconstruction Loss. We define the distance between our reconstructed graph and the original graph $\|\hat{\mathbf{G}} - \mathbf{G}\|$. It includes multiple components to ensure the fidelity of generated 3D molecules:

$$\begin{aligned}\mathcal{L}_{\text{atom}} &= \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \text{CE}(\hat{\mathbf{F}}_i, \mathbf{F}_i), & \mathcal{L}_{\text{coordinate}} &= \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \text{MSE}(\hat{\mathbf{X}}_i, \mathbf{X}_i), \\ \mathcal{L}_{\text{bond}} &= \frac{1}{|\mathcal{V}|^2} \sum_{i,j \in \mathcal{V}} \text{CE}(\hat{\mathbf{E}}_{ij}, \mathbf{E}_{ij}), & \mathcal{L}_{\text{distance}} &= \frac{1}{|\mathcal{V}|^2} \sum_{i,j \in \mathcal{V}} w_{ij} \text{MSE}(\|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}_j\|, \mathbf{D}_{ij}).\end{aligned}\quad (10)$$

where CE denotes cross-entropy loss, and $\mathcal{L}_{\text{atom}}$ and $\mathcal{L}_{\text{bond}}$ are losses for atom and bond types; $\mathcal{L}_{\text{coordinate}}$ measures the MSE between the predicted and the ground truth coordinates; and $\mathcal{L}_{\text{distance}}$ serves as an extra constraint on coordinate predictions, to ensure correct inter-atomic distances. Notably, we incorporate a *bonded distance loss* inside $\mathcal{L}_{\text{distance}}$ by adjusting the pairwise weight w_{ij} :

$$w_{ij} = \begin{cases} 1 + \lambda, & \text{if atom } i \text{ and } j \text{ are bonded,} \\ 1, & \text{otherwise.} \end{cases} \quad (11)$$

This design prioritizes accurate inter-atomic distances between bonded atoms, enforcing stricter geometry in chemically important regions. λ is a hyperparameter that controls this constraint’s strength. Given the components above, we define the distance between our reconstructed graph and the original graph (*i.e.*, the reconstruction loss) as follows:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathbf{G} \sim p_{\text{data}}} \|\hat{\mathbf{G}} - \mathbf{G}\| = \mathbb{E}_{\mathbf{G} \sim p_{\text{data}}} \boldsymbol{\gamma} \cdot [\mathcal{L}_{\text{atom}}, \mathcal{L}_{\text{bond}}, \mathcal{L}_{\text{coordinate}}, \mathcal{L}_{\text{distance}}]^\top, \quad (12)$$

where $\boldsymbol{\gamma} \in \mathbb{R}^4$ is a hyperparameter vector balancing the reconstruction terms. The final loss for UAE-3D combines $\mathcal{L}_{\text{recon}}$ with a KL regularization term:

$$\mathcal{L}_{\text{UAE-3D}} = \mathcal{L}_{\text{recon}} + \beta \cdot \mathbb{E}_{\mathbf{G} \sim p_{\text{data}}} [D_{\text{KL}}(q(\mathbf{Z}|\mathbf{G}) \| p(\mathbf{Z}))], \quad (13)$$

where the hyperparameter $\beta \in \mathbb{R}$ controls the KL regularization strength.

SE(3)-Equivariant Augmentations. To enforce SE(3)-equivariance in UAE-3D, we apply random transformations $\mathbf{R} \in \text{SE}(3)$ to input coordinates \mathbf{X} during training [28, 20]. Each \mathbf{R} combines a rotation from $\text{SO}(3)$ and a random translation from $\mathcal{N}(\mathbf{0}, 0.01\mathbf{I}_3)$. Crucially, the reconstruction loss $\mathcal{L}_{\text{recon}}$ (*cf.* Equation (12)) remains unchanged but operates on transformed inputs $\mathbf{R} \circ \mathbf{G} = \langle \mathbf{F}, \mathbf{E}, \mathbf{R}(\mathbf{X}) \rangle$. This process can be informally conceptualized as a reconstruction loss of $\|\mathcal{D}(\mathcal{E}(\mathbf{R} \circ \mathbf{G})) - \mathbf{R} \circ \mathbf{G}\|$, where the autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$ learns to preserve geometric consistency under SE(3) transformations.

3.2 Unified Latent Diffusion Modeling for 3D Molecule Generation

Diffusion Transformer (DiT). Given the unified latent space provided by our UAE-3D model, we adopt the DiT [23] as the backbone diffusion model ϵ_θ for 3D molecular latent generation. Originally developed for image LDM, DiT has demonstrated strong performance in modeling latent sequences. Compared to a standard Transformer [22], DiT replaces the layernorm [29] by adaptive layernorm (adaLN) [30], where the scale $\mathbf{y} \in \mathbb{R}^d$ and shift $\mathbf{b} \in \mathbb{R}^d$ parameters are conditioned on the diffusion timestep $t \in (0, 1]$. Specifically, $[\mathbf{y}; \mathbf{b}] = \text{MLP}_{\text{adaLN}}(\text{Embed}_t(t))$, where $\text{Embed}_t(\cdot)$ is a shared linear layer and $\text{MLP}_{\text{adaLN}}$ is a module specific to each adaLN instance. These generated shift and scale parameters act as “soft gates”, enabling timestep-dependent activation of DiT’s hidden representations and facilitating more effective, scale-adaptive denoising.

Conditional Generation with Classifier-Free Guidance. To enable generation with a condition vector \mathbf{c} , we extend the adaLN modules by combining the condition embedding $\text{Embed}_c(\mathbf{c})$ with the timestep embedding $\text{Embed}_t(t)$ when computing the scale \mathbf{y} and shift \mathbf{b} parameters. This allows the diffusion model $\epsilon_\theta(\mathbf{Z}^{(t)}, t, \mathbf{c})$ to incorporate conditioning information during denoising:

$$[\mathbf{y}; \mathbf{b}] = \text{MLP}_{\text{adaLN}}(\text{Embed}_t(t) + \text{Embed}_c(\mathbf{c})). \quad (14)$$

To further enforce conditioning during inference, we employ the classifier-free guidance (CFG) [31] to find \mathbf{Z} that has a high $\log p(\mathbf{c}|\mathbf{Z})$. By Bayes’ rule, the gradient of this objective is $\nabla_{\mathbf{Z}} \log p(\mathbf{c}|\mathbf{Z}) \propto \nabla_{\mathbf{Z}} \log p(\mathbf{Z}|\mathbf{c}) - \nabla_{\mathbf{Z}} \log p(\mathbf{Z})$. Following [32], we can interpret the diffusion model ϵ_θ ’s output as score functions, and therefore maximize $p(\mathbf{c}|\mathbf{Z})$ using the modified denoising function:

$$\tilde{\epsilon}_\theta(\mathbf{Z}^{(t)}, t, \mathbf{c}) = (1 + w)\epsilon_\theta(\mathbf{Z}^{(t)}, t, \mathbf{c}) - w\epsilon_\theta(\mathbf{Z}^{(t)}, t), \quad (15)$$

Table 1: Performance of *de novo* 3D molecule generation on GEOM-Drugs. * indicates results reproduced using official source codes, while other baseline results are taken from [6].

2D-Metric	FCD \downarrow	AtomStable	MolStable	V&C	V&U	V&U&N	SNN	Frag	Scaf
Train	0.251	1.000	1.000	1.000	1.000	0.000	0.585	0.999	0.584
CDGS	22.051	0.991	0.706	0.285	0.285	0.285	0.262	0.789	0.022
JODO	2.523	1.000	0.981	0.874	0.905	0.902	0.417	0.993	0.483
MiDi*	7.054	0.968	0.822	0.633	0.654	0.652	0.392	0.951	0.196
EQGAT-diff*	5.898	1.000	0.989	0.845	0.863	0.859	0.377	0.983	0.161
UDM-3D, ours	0.692 <small>−72.6%</small>	1.000	0.925	0.879	0.913	0.907	0.525	0.990	0.540
3D-Metric	FCD _{3D} \downarrow	AtomStable	MolStable	Bond length \downarrow	Bond angle \downarrow	Dihedral angle \downarrow			
Train	13.73	0.861	0.028	1.56E-04	1.81E-04		1.56E-04		
EDM	31.29	0.831	0.002	4.29E-01	4.96E-01		1.46E-02		
JODO	19.99	0.845	0.010	8.49E-02	1.15E-02		6.68E-04		
MiDi*	23.14	0.750	0.003	1.17E-01	9.57E-02		4.46E-03		
GeoLDM	30.68	0.843	0.008	3.91E-01	4.22E-01		1.69E-02		
EQGAT-diff*	26.33	0.825	0.007	1.55E-01	5.21E-02		2.10E-03		
UDM-3D, ours	17.36 <small>−13.2%</small>	0.852	0.014	9.89E-03 <small>−88.4%</small>	5.11E-03 <small>−55.6%</small>		1.74E-04 <small>−74.0%</small>		

where $w \in [0, +\infty)$ is a hyperparameter controlling the guidance strength; $\epsilon_\theta(\mathbf{Z}^{(t)}, \mathbf{t}, \mathbf{c})$ is the conditional variant of ϵ_θ , incorporating the property \mathbf{c} ; $\epsilon_\theta(\mathbf{Z}^{(t)}, \mathbf{t})$ is the unconditioned variant that ignores \mathbf{c} . To train ϵ_θ , we randomly drop condition \mathbf{c} with a certain probability p_{drop} , allowing ϵ_θ to learn both conditional and unconditional distributions.

4 Experiments

In this section, we present the experimental results of our proposed unified molecular latent space approach (UAE-3D & UDM-3D) on 3D molecule generation tasks. We comprehensively evaluate UDM-3D’s performance on *de novo* 3D molecule generation and conditional 3D molecule generation with targeted quantum properties. Our experiments demonstrate significant improvements over state-of-the-art methods while maintaining computational efficiency.

4.1 Experimental Setup

Datasets. Following prior works [5, 6, 9], we conduct experiments on QM9 [24] and GEOM-Drugs [25] datasets. QM9 [24] contains 130k small organic molecules (up to 9 heavy atoms) with quantum chemical properties. We use 100K/18K/13K train/val/test splits, following [5, 6]. GEOM-Drugs [25] is a pharmaceutical-scale dataset with 450k drug-like molecules (average 44.4 atoms, max 181 atoms). We use 80%/10%/10% splits and retain the lowest-energy conformer per molecule.

Baselines. For *de novo* 3D molecule generation, we compare UDM-3D with CDGS [33], JODO [6], MiDi [9], G-SchNet [12], G-SphereNet [34], EDM [5], MDM [35], GeoLDM [17], EQGAT-diff [36], SemlaFlow[37] and ADiT [38]. For conditional generation, we additionally use baselines of EEGSDE [8] and GeoBFN [39]. We do not report MDM’s performance on GEOM-DRUGS because we are unable to reproduce reasonable results using their official code. Similarly, ADiT’s performance on GEOM-DRUGS is omitted, because it is not available in their paper, code, and released checkpoint.

4.2 De Novo 3D Molecule Generation

Experimental Setting. Generating a 3D molecule involves generating the structural validity (atom/bond features) and accurate spatial arrangements (3D coordinates). Therefore, we adapt the comprehensive metrics from [6, 5]. We also report all the metrics for molecules from training set, serving as an approximate upper bound for performance. Our evaluation metrics can be divided into two groups: (1) **2D Metrics**: Atom stability, validity & completeness (V&C), validity & uniqueness (V&U), validity & uniqueness & novelty (V&U&N), similarity to nearest neighbor (SNN), fragment similarity (Frag), scaffold similarity (Scaf), and Fréchet ChemNet Distance (FCD) [3]; (2) **3D Metrics**: Atom stability, FCD, maximum mean discrepancy (MMD) [40] for the distributions of bond lengths, bond angles, and dihedral angles. More experimental details are in Appendix C.1.

Table 1 and Table 2 present the performance of UDM-3D on the *de novo* generation task on QM9 and GEOM-Drugs datasets, respectively. Our model demonstrates leading performances in generat-

Table 2: Performance of *de novo* 3D molecule generation on QM9. * indicates results reproduced using official source codes, while other baseline results are taken from [6]. Some of ADiT’s evaluation metrics are omitted due to differences in evaluation protocols. We discuss these protocol differences and their implications in Appendix C.

2D-Metric	FCD \downarrow	AtomStable	MolStable	V&C	V&U	V&U&N	SNN	Frag	Scaf
Train	0.063	0.999	0.988	0.989	0.989	0.000	0.490	0.992	0.946
CDGS	0.798	0.997	0.951	0.951	0.936	0.860*	0.493	0.973	0.784
JODO	0.138	0.999	0.988	0.990	0.960	0.780*	0.522	0.986	0.934
MiDi*	0.187	0.998	0.976	0.980	0.954	0.769	0.501	0.979	0.882
EQGAT-diff*	2.088	0.999	0.971	0.965	0.950	0.891	0.482	0.950	0.703
SemlaFlow	0.863	0.995	0.949	0.857	0.821	0.821	0.124	-	-
UDM-3D, ours	0.130 <small>-5.80%</small>	0.999	0.988	0.983	0.973	0.950	0.508	0.987	0.898
3D-Metric	FCD _{3D} \downarrow	AtomStable	MolStable	Bond length \downarrow	Bond angle \downarrow	Dihedral angle \downarrow			
Train	0.877	0.994	0.953	5.44E-04	4.65E-04	1.78E-04			
EDM	1.285	0.986	0.817	1.30E-01	1.82E-02	6.64E-04			
MDM	4.861	0.992	0.896	2.74E-01	6.60E-02	2.39E-02			
JODO	0.885	0.992	0.934	1.48E-01	1.21E-02	6.29E-04			
GeoLDM	1.030	0.989	0.897	2.40E-01	1.00E-02	6.59E-04			
MiDi*	1.100	0.983	0.842	8.96E-01	2.08E-02	8.14E-04			
EQGAT-diff*	1.520	0.988	0.888	4.21E-01	1.89E-02	1.24E-03			
SemlaFlow	1.127	0.971	0.787	-	-	-			
ADiT*	2.884	0.211	-	9.98E-01	3.38E-02	1.46E-03			
UDM-3D, ours	0.881 <small>-0.45%</small>	0.993	0.935	7.04E-02 <small>-45.8%</small>	9.84E-03 <small>-1.6%</small>	3.47E-04 <small>-44.8%</small>			

ing chemically valid and novel molecules and achieves state-of-the-art performance in most of the metrics across both datasets. Two detailed key observations emerge from the results:

Leading Geometric Accuracy. Our unified latent space enables highly accurate 3D geometric modeling, a key factor in realistic molecule generation. UDM-3D significantly outperforms baselines in distributional distances of bond lengths and angles, reducing errors by an order of magnitude. These improvements hold even for complex, drug-like molecules in GEOM-Drugs where UDM-3D achieves a **25x** lower bond length error than GeoLDM (9.89E-03 *v.s.* 3.91E-01). This leap in accuracy comes from UDM-3D’s ability to jointly optimize chemical and geometric constraints in a unified latent space, maintaining consistency across molecular modalities.

Novel and High Quality Generation. UDM-3D achieves outstanding V&U&N scores of **0.907** on GEOM-Drugs and **0.950** on QM9, while achieving state-of-the-art 3D stability and comparable 2D stability to strong baselines such as JODO. This highlights its ability to generate novel molecules without compromising structural quality. The key lies in the unified latent space, which jointly encodes atom types, bond types, and coordinates, enabling the model to capture complex interdependencies among molecular features.

4.3 Conditional 3D Molecule Generation

Experimental Settings. We evaluate conditional generation of molecules with target quantum properties using the protocol from [5, 17]. Specifically, our target properties include C_v , μ , α , ϵ_{HOMO} , ϵ_{LUMO} , $\Delta\epsilon$, and we evaluate the Mean Absolute Error (MAE) between target and predicted properties. More details on the properties and settings are provided in Appendix C.2.

Table 3 reports MAE results for conditional generation on QM9. UDM-3D achieves the lowest MAE in five out of six properties, showing strong capability in generating molecules that match target properties. On average, it reduces MAE by 42.2% compared to GeoLDM, with a notable 52.7% improvement in predicting the HOMO-LUMO gap ($\Delta\epsilon$). These results highlight the importance of our unified latent space that effectively integrates all molecular modalities during generation. While GeoLDM struggles to correlate all the modalities due to its fragmented latent spaces, our unified latent representation inherently preserves the interplay between 3D geometry, bonds, and electronic characteristics, enabling precise modeling over molecule-property relationships.

4.4 Training & Sampling Efficiency

We further analyze the training and sampling efficiency of UDM-3D compared to baselines on the QM9 dataset. These experiments are performed with an NVIDIA A100 GPU.

Table 3: Mean Absolute Error (MAE) for conditional 3D molecule generation on QM9.

Method	μ (D)	α (Bohr ³)	C_v (cal/mol K)	$\varepsilon_{\text{HOMO}}$ (meV)	$\varepsilon_{\text{LUMO}}$ (meV)	$\Delta\varepsilon$ (meV)
U-Bound	1.613	8.98	6.879	645	1457	1464
L-Bound	0.043	0.09	0.040	39	36	65
EDM	1.123	2.78	1.065	371	601	671
EEGSDE	0.777	2.50	0.941	302	447	487
GeoLDM	1.108	2.37	1.025	340	522	587
GeoBFN	0.998	2.34	0.949	328	516	577
JODO	0.628	1.42	0.581	226	256	335
UDM-3D, ours	0.603 <small>-3.98%</small>	1.54 <small>+8.4%</small>	0.553 <small>-4.82%</small>	216 <small>-4.42%</small>	247 <small>-3.5%</small>	313 <small>-6.57%</small>

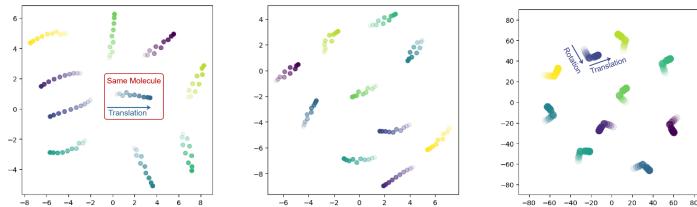


Figure 4: t-SNE visualizations of UAE-3D’s latents under SE(3) augmentations. (a) Translations along a fixed direction. (b) Rotations along a fixed axis. (c) Sequential rotations followed by translations. Color gradients show increasing distances or angles.

Table 4: Comparing DiT with Transformer and PerceiverIO for latent diffusion modeling. Models have the same depth and hidden size. A.S. → atom stability.

Metric	DiT	Transformer	PerceiverIO
A.S. _{3D}	0.993	0.983	0.972
A.S. _{2D}	0.999	0.997	0.990
V&C	0.983	0.938	0.933
V&U	0.973	0.922	0.931
V&U&N	0.950	0.922	0.931

Training Efficiency. As Figure 2(c) shows, UDM-3D’s training costs 52 hours (14h UAE-3D + 38h UDM-3D), which is 5.3 times faster than GeoLDM and 2.7 times faster than JODO. This efficiency stems from: (1) Our decoupled training paradigm: UAE-3D first learns 3D molecule compression, allowing DiT to focus solely on generative modeling of the compressed latents. (2) Efficient diffusion with DiT: Unlike previous molecular diffusion models that require complex architectures for multi-modality and equivariance, DiT’s simple and highly parallelizable design allows for faster training and sampling in the unified latent space. Crucially, our speedup does not compromise performance, as guaranteed by UAE-3D’s near-lossless molecular reconstruction (cf. Figure 2(a;b)).

Sampling Speed. Figure 2(d) shows that UDM-3D generates each molecule in just 0.081s, which is 7.3x faster than EDM/GeoLDM and 9.8x faster than JODO. This speed advantage comes from our unified latent space, which simplifies DiT’s modeling and avoids complex neural architectures.

4.5 Analysis and Ablation Studies

UAE-3D’s Latents Present Structured Variations across Geometric Movements. Figure 4 shows the t-SNE visualizations of UAE-3D’s latent representation under different SE(3) augmentations. Each dot represents a molecular embedding in the latent space, and the color gradient from light to dark corresponds to an increase in the translation distance (middle) or the rotation angle (right) for the same molecule. We observe that the same molecule’s representations remain close in the latent space and as the augmentation scale increases, the representation changes along a consistent direction. Figure 4c shows that the representations change moving directions when the augmentation is switched from rotation to translation. These observations demonstrate that UAE-3D captures meaningful and structured variations in molecular representations across SE(3) augmentations.

3D Molecule Samples by UDM-3D. We present 3D molecules generated by our method as case studies in Figure 5. The generated molecules are chemically valid and exhibit diverse and complex structures, showcasing UDM-3D’s ability to generate realistic 3D molecular conformations.

Ablating the DiT Backbone. We compare DiT with alternative transformer-based architectures for *de novo* 3D molecule generation on the QM9 dataset (cf. Table 4). Specifically, we evaluate against a vanilla Transformer and PerceiverIO[41], a modern architecture designed for structured inputs and outputs. The results show that DiT consistently outperforms both baselines across all metrics, which we attribute to its adaptive LayerNorm layers. These layers enable DiT to effectively handle data with varying noise scales, thereby improving diffusion performance. Moreover, when paired with alternative diffusion neural architectures such as Transformer or PerceiverIO, our UAE-3D framework still achieves meaningful and comparable performances, demonstrating its robustness

Table 5: The performance influence of SE(3)-equivariant augmentations on the QM9 dataset.

2D-Metric	FCD \downarrow	AtomStable	MolStable	V&C	V&U	V&U&N	SNN	Frag	Scaf
No aug.	0.581	0.995	0.947	0.950	0.935	0.921	0.492	0.971	0.799
+ Rot.	0.315	0.995	0.948	0.951	0.943	0.927	0.493	0.979	0.875
+ Trans.	0.202	0.999	0.986	0.980	0.967	0.944	0.507	0.981	0.884
+ Trans. + Rot.	0.130	0.999	0.988	0.983	0.973	0.950	0.508	0.987	0.898
3D-Metric	FCD _{3D} \downarrow	AtomStable	MolStable	Bond length \downarrow	Bond angle \downarrow	Dihedral angle \downarrow			
No aug.	1.065	0.993	0.879	1.01E-01	1.18E-02		1.08E-03		
+ Rot.	1.007	0.993	0.876	8.12E-02	9.99E-03		3.90E-04		
+ Trans.	0.902	0.989	0.896	7.08E-02	1.04E-02		6.81E-04		
+ Trans. + Rot.	0.881	0.993	0.935	7.04E-02	9.84E-03		3.47E-04		

Table 6: RMSD results when trained with different SE(3) augmentations on QM9.

Train with	RMSD ($\times 10^{-3}$ Å)
No aug.	1.4
+ Rot.	1.1
+ Trans.	0.9
+ Trans. + Rot.	0.2

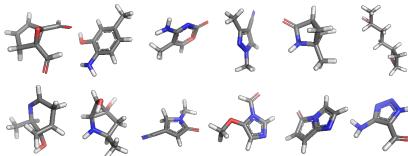


Figure 5: Visualization of random samples generated by UDM-3D on QM9.

to different backbone choices. While further hyperparameter tuning could potentially improve the performance of these alternatives, the results already highlight the advantage of DiT in our setting.

SE(3)-Equivariant Augmentations. Table 5 shows the performance influence of various SE(3)-equivariant augmentations on the QM9 dataset. The results reveal that data augmentation improves overall performance by providing diverse geometric views during UDM-3D training.

In addition to the standard performance metrics, we further investigate UAE-3D’s reconstruction error on the test set when trained under different 3D transformations. Table 6 presents the RMSD of the reconstructed molecules. By including both rotation and translation augmentations, we observe a significant reduction in reconstruction error (from 1.4×10^{-3} to 0.2×10^{-3}), achieving the lowest RMSD. This shows that our SE(3)-equivariant augmentations effectively enhance the model’s ability to learn geometric consistency by exploring a more comprehensive 3D molecular space.

5 Related Works

3D Molecule Generation. Early efforts for 3D molecule generation focus on autoregressive approaches [12, 13, 34], constructing 3D molecules sequentially by adding atoms or molecular fragments. With the success of diffusion models across various domains [42, 15, 43], they also become the *de facto* method for 3D molecule generation [5, 8]. However, the early diffusion works can easily generate invalid molecules because of overlooking the bond information. To bridge this gap, subsequent works [6, 9, 11] additionally consider bond generation by building a separate diffusion process. However, generating different modalities in separate diffusion processes unnecessarily complicates the model design, reducing efficiency. It also risks compromising the consistency between these modalities. To address this, our UDM-3D performs generative modeling in UAE-3D’s unified latent space that integrates all molecular modalities, improving efficiency and generation quality.

LDMs for 3D Molecule Generation. For efficiency, LDMs employ a VAE [16] to compress raw data into a low-dimensional latent space, where a diffusion model performs generative modeling [14]. This method has been very popular for image generation [23]. However, existing 3D molecular LDMs still face challenges in separated latent spaces. For example, GeoLDM [17] compresses atom features and 3D coordinates separately, failing to build a unified latent space. Other works [39, 44] process different molecular components (atoms/subgraphs or atom types/coordinates) in separate channels, increasing modeling complexity. In 3D protein generation, PepGLAD [18] compresses 1D sequences and 3D structures into two separate latent spaces. Our UDM-3D addresses these issues by employing a unified latent space, significantly improved generation efficiency.

6 Conclusion and Future Works

In this paper, we propose UAE-3D to compress the multi-modal features of 3D molecules into a unified latent space, and demonstrate the effectiveness of latent diffusion modeling on this space by introducing UDM-3D. By integrating atom types, chemical bonds, and 3D coordinates into a single latent space, our model effectively addresses the inherent challenges of multi-modality and $SE(3)$ equivariance for 3D molecule generation. Extensive experiments on GEOM-Drugs and QM9 confirm leading performances in both *de novo* and conditional 3D molecule generation, setting new benchmarks for both quality and efficiency. UAE-3D’s success highlights the benefits of building a unified latent space for molecular design. Moving forward, we will transfer this unified latent space to new modalities, including proteins and RNAs, and extend it to broader molecular design tasks, including structure-based drug design. While UAE-3D’s latents are near-lossless compressions of the original molecule, it inspires the exploration of jointly modeling UAE-3D’s latents with text sequences for text-guided 3D molecule generation, following [45–51], as well as joint modeling of molecular and biological text representations across broader domains, including proteins [52] and single-cell transcriptomics [53].

Limitations

Scope. In this work, we explore 3D molecule generation under the settings of unconditional generation and conditional generation targeting at quantum chemical properties. Other molecule generation tasks, such as structure-based drug design [54], is out of the scope of this work. We leave this exploration to future works.

Molecule Size. Our method relies on a separate module to decide each generated molecule’s number of atom, following prior works [6, 5, 8]. Specifically, we use the molecule size distribution measured on the training dataset to sample new molecule size. While this method shows decent performance on the tested tasks, we conjecture this will be one of the bottlenecks when the tested task become more complex. Resolving this issue might demand the incorporation of auto-regressive based method for molecule generation, in which the generated molecule size is automatically controlled by the auto-regressive process.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-002), the Ministry of Education (MOE T1251RES2309 and MOE T2EP20125-0039), the Agency for Science, Technology and Research (A*STAR) (H25J6a0034), and the National Natural Science Foundation of China (62572449).

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

References

- [1] Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2328–2337. PMLR, 2018.
- [2] Ross Irwin, Spyridon Dimitriadis, Jiazen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.*, 3(1):15022, 2022.
- [3] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Ok-tai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 2020.
- [4] Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. In *ICLR*. OpenReview.net, 2020.
- [5] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 8867–8887. PMLR, 2022.
- [6] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2-d and 3-d graph diffusion models for complete molecule generation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [7] Zaixi Zhang, Yaosen Min, Shuxin Zheng, and Qi Liu. Molecule generation for target protein binding with structural motifs. In *ICLR*. OpenReview.net, 2023.
- [8] Fan Bao, Min Zhao, Zhongkai Hao, Peiyao Li, Chongxuan Li, and Jun Zhu. Equivariant energy-guided SDE for inverse molecular design. In *ICLR*. OpenReview.net, 2023.
- [9] Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *ECML/PKDD (2)*, volume 14170 of *Lecture Notes in Computer Science*, pages 560–576. Springer, 2023.
- [10] Tuan Le, Julian Cremer, Frank Noé, Djork-Arné Clevert, and Kristof T. Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *ICLR*. OpenReview.net, 2024.
- [11] Chenqing Hua, Sitao Luan, Minkai Xu, Zhitao Ying, Jie Fu, Stefano Ermon, and Doina Precup. Mudiff: Unified diffusion for complete molecule generation. In *LoG*, volume 231 of *Proceedings of Machine Learning Research*, page 33. PMLR, 2023.
- [12] Niklas W. A. Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *NeurIPS*, pages 7564–7576, 2019.
- [13] 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. *CoRR*, abs/2109.04824, 2021.
- [14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022.
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.
- [16] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[17] Minkai Xu, Alexander S. Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 38592–38610. PMLR, 2023.

[18] Xiangzhe Kong, Yinjun Jia, Wenbing Huang, and Yang Liu. Full-atom peptide design with geometric latent diffusion. In *NeurIPS*, 2024.

[19] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 2021.

[20] Yuyang Wang, Ahmed A. A. Elhag, Navdeep Jaitly, Joshua M. Susskind, and Miguel Ángel Bautista. Swallowing the bitter pill: Simplified scalable conformer generation. In *ICML*. OpenReview.net, 2024.

[21] Cameron Diao and Ricky Loynd. Relational attention: Generalizing transformers for graph-structured tasks. In *ICLR*. OpenReview.net, 2023.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4172–4182. IEEE, 2023.

[24] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 2014.

[25] Simon Axelrod and Rafael Gómez-Bombarelli. GEOM: energy-annotated molecular conformations for property prediction and molecular generation. *CoRR*, abs/2006.05531, 2020.

[26] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *CoRR*, abs/2107.00630, 2021.

[27] Philipp Thölke and Gianni De Fabritiis. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022.

[28] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[29] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[30] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, pages 3942–3951. AAAI Press, 2018.

[31] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*. OpenReview.net, 2021.

[33] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Conditional diffusion based on discrete graph structures for molecular graph generation. In *AAAI*, pages 4302–4311. AAAI Press, 2023.

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

[35] Lei Huang, Hengtong Zhang, Tingyang Xu, and Ka-Chun Wong. MDM: molecular diffusion model for 3d molecule generation. In *AAAI*, pages 5105–5112. AAAI Press, 2023.

[36] Tuan Le, Frank Noé, and Djork-Arné Clevert. Representation learning on biomolecular structures using equivariant graph attention. In *LoG*, volume 198 of *Proceedings of Machine Learning Research*, page 30. PMLR, 2022.

[37] Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Semlaflow - efficient 3d molecular generation with latent attention and equivariant flow matching. In *AISTATS*, volume 258 of *Proceedings of Machine Learning Research*, pages 3772–3780. PMLR, 2025.

[38] Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*, 2025.

[39] Yuxuan Song, Jingjing Gong, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified generative modeling of 3d molecules with bayesian flow networks. In *The Twelfth International Conference on Learning Representations*, 2023.

[40] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[41] Andrew Jaegle, Sébastien Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *ICLR*. OpenReview.net, 2022.

[42] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015.

[43] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021.

[44] Tian Bian, Yifan Niu, Heng Chang, Divin Yan, Junzhou Huang, Yu Rong, Tingyang Xu, Jia Li, and Hong Cheng. Hierarchical graph latent diffusion model for conditional molecule generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 130–140. ACM, 2024.

[45] Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *EMNLP*, pages 375–413. Association for Computational Linguistics, 2022.

[46] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *EMNLP*, 2023.

[47] Zhiyuan Liu, Yanchen Luo, Han Huang, Enzhi Zhang, Sihang Li, Junfeng Fang, Yaorui Shi, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. NEXT-MOL: 3d diffusion meets 1d language modeling for 3d molecule generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[48] Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 3d-molm: Towards 3d molecule-text interpretation in language models. In *ICLR*, 2024.

[49] Yanchen Luo, Junfeng Fang, Sihang Li, Zhiyuan Liu, Jiancan Wu, An Zhang, Wenjie Du, and Xiang Wang. Text-guided small molecule generation via diffusion model. *iScience*, 27(11), 2024.

[50] Zhiyuan Liu, Yaorui Shi, An Zhang, Sihang Li, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. Reactxt: Understanding molecular reaction-ship via reaction-contextualized molecule-text pretraining. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024.

[51] Liang Wang, Chao Song, Zhiyuan Liu, Yu Rong, Qiang Liu, Shu Wu, and Liang Wang. Diffusion models for molecules: A survey of methods and tasks. *arXiv*, abs/2502.09511, 2025.

[52] Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. Prott3: Protein-to-text generation for text-based protein understanding. In *ACL*. Association for Computational Linguistics, 2024.

[53] Yaorui Shi, Jiaqi Yang, Changhao Nai, Sihang Li, Junfeng Fang, Xiang Wang, Zhiyuan Liu, and Yang Zhang. Language-enhanced representation learning for single-cell transcriptomics. *arXiv preprint arXiv:2503.09427*, 2025.

[54] Yanru Qu, Keyue Qiu, Yuxuan Song, Jingjing Gong, Jiawei Han, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Molcraft: Structure-based drug design in continuous parameter space. *arXiv preprint arXiv:2404.12141*, 2024.

[55] Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *ICANN (1)*, volume 11139 of *Lecture Notes in Computer Science*, pages 412–422. Springer, 2018.

[56] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, pages 6306–6315, 2017.

[57] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, pages 14837–14847, 2019.

[58] Myeonghun Lee and Kyoungmin Min. MGCVAE: multi-objective inverse design via molecular graph conditional variational autoencoder. *J. Chem. Inf. Model.*, 62(12):2943–2950, 2022.

[59] Seokho Kang and Kyunghyun Cho. Conditional molecular design with deep generative models. *J. Chem. Inf. Model.*, 59(1):43–52, 2019.

[60] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L. Gaunt. Constrained graph variational autoencoders for molecule design. In *NeurIPS*, pages 7806–7815, 2018.

[61] Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. Molgensurvey: A systematic survey in machine learning models for molecule design. *CoRR*, abs/2203.14500, 2022.

[62] Bin Dai and David P. Wipf. Diagnosing and enhancing VAE models. In *ICLR (Poster)*. Open-Review.net, 2019.

[63] Jyoti Aneja, Alexander G. Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. In *NeurIPS*, pages 480–493, 2021.

[64] Zhiyuan Liu, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Rethinking tokenizer and decoder in masked graph modeling for molecules. In *NeurIPS*, 2023.

[65] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *NeurIPS*, pages 11287–11302, 2021.

[66] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: latent point diffusion models for 3d shape generation. In *NeurIPS*, 2022.

[67] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

- [68] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008.
- [69] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. PMID: 8709122.
- [70] Jan Jensen. Xyz2mol.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Our claims in abstract and introduction are well supported by results in the experimental section.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitation section is Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The detailed needed for reproducibility are included in Section 4, Appendix C, and our released code at <https://github.com/lyc0930/UAE-3D/>.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: **[Yes]**

Justification: Our code and used dataset is released at <https://github.com/lyc0930/UAE-3D/>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: **[Yes]**

Justification: The settings are detailed in Section 4 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: **[No]**

Justification: The experiments are mostly computationally expensive and we did not report error bars, following the baseline methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: These information are included in Section 4 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: We adhere to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[NA\]](#)

Justification: This work focuses on generating small molecules. It is pure computational research for applications in chemistry, material, and biology without possible harm to the human society

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our method does not involve such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The datasets are cited in Section 4 and Appendix C.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our codes released at <https://github.com/lyc0930/UAE-3D/> include a readme file as documentation.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We have no crowdsourcing and no involvement of human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorosity, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A More Experimental Results

A.1 More Ablation Studies and Analysis

Table 7: The influence of latent space dimensionality for 3D molecule generation on the QM9 dataset.

Latent Dim	Atom Acc	Bond Acc	RMSD	AtomStable	MolStable	V&C	V&U	V&U&N	Bond Length↓	Bond Angle↓	Dihedral Angle↓
4	0.9201	0.8932	0.0080	0.914	0.325	0.580	0.566	0.566	3.27E-01	2.20E-01	4.19E-03
8	0.9998	0.9746	0.0006	0.987	0.882	0.942	0.923	0.923	1.29E-01	1.77E-02	8.14E-04
16	1.0000	1.0000	0.0002	0.999	0.988	0.983	0.973	0.950	7.04E-02	9.84E-03	3.47E-04
32	1.0000	1.0000	0.0003	0.986	0.867	0.943	0.918	0.918	9.98E-02	1.42E-02	7.77E-04

Table 8: The influence of R-Trans layer depth for reconstruction on the QM9 dataset.

R-Trans layers	Atom Acc	Bond Acc	RMSD
3	1.0000	0.9999	0.000326
6	1.0000	1.0000	0.000203
9	1.0000	1.0000	0.000229
12	1.0000	1.0000	0.000209

Table 9: The influence of different bonded distance loss weight λ on the QM9 dataset.

λ / VAE Recon.		Atom Accuracy (%)			Bond Accuracy (%)			Coordinate RMSD (Å)		
0 (w/o λ)		1.000			1.000			0.0034		
1		1.000			1.000			0.0010		
5		1.000			1.000			0.0007		
10		1.000			1.000			0.0002		
20		1.000			1.000			0.0002		
λ / 2D-Metric		FCD↓	AtomStable	MolStable	V&C	V&U	V&U&N	SNN	Frag	Scaf
1		0.283	0.996	0.959	0.962	0.940	0.940	0.494	0.978	0.874
5		0.255	0.997	0.965	0.970	0.947	0.947	0.500	0.983	0.888
10		0.130	0.999	0.988	0.983	0.973	0.950	0.508	0.987	0.898
20		0.172	0.999	0.988	0.980	0.969	0.947	0.507	0.987	0.898
λ / 3D-Metric		FCD _{3D} ↓	AtomStable	MolStable	Bond length↓		Bond angle↓	Dihedral angle↓		
1		0.901	0.985	0.857	1.25E-01		1.80E-02	1.06E-03		
5		0.892	0.989	0.909	1.12E-01		1.19E-02	7.18E-04		
10		0.881	0.993	0.935	7.04E-02		9.84E-03	3.47E-04		
20		0.889	0.993	0.929	7.05E-02		9.84E-03	3.45E-04		

Reconstruction Errors. To assess the reconstruction fidelity of UAE-3D, we evaluate its performance on the held-out test sets of QM9 and GEOM-Drugs. We report atom-type and bond-type accuracies as well as the coordinates RMSD, which jointly reflect the VAE’s ability to compress and reconstruct both invariant and equivariant molecular features. Table 11 presents the reconstruction results. On the QM9 dataset, UDM-3D achieves perfect reconstruction accuracy (100% for both atoms and bonds) with a near-zero coordinate RMSD of 0.0002 Å, substantially overperforming GeoLDM and ADiT. Notably, ADiT does not retain and reconstruct chemical bond information, which is a limitation when reconstructing 3D molecules. On the more complex GEOM-Drugs dataset, UDM-3D again achieves perfect reconstruction (100.0% atom and bond accuracy, 0.0008 Å RMSD), whereas GeoLDM exhibits noticeable degradation. This demonstrates that UDM-3D’s unified latent space not only faithfully compresses all molecular modalities but also generalizes robustly to large and diverse molecular datasets. These results show that UAE-3D effectively learns a unified latent representation that preserves both the chemical identity and geometric precision of molecules. These findings align with the visual comparison in Figure 2(a;b), and reinforce our claim that UAE-3D achieves near-lossless reconstruction an essential prerequisite for accurate latent diffusion modeling.

Unified vs. Separated Latent Space. To validate the effectiveness of our unified latent space design, we conduct an ablation study comparing UDM-3D with a variant that employs separate latent spaces for invariant (2D) and equivariant (3D) molecular features. Specifically, we split the unified VAE

Table 10: Comparison between unified and separated latent space on the QM9 dataset.

2D-Metric	FCD \downarrow	AtomStable	MolStable	V&C	V&U	V&U&N	SNN	Frag	Scaf
Unified	0.130	0.999	0.988	0.983	0.973	0.950	0.508	0.987	0.898
	0.351	0.995	0.952	0.943	0.920	0.913	0.341	0.940	0.682
3D-Metric	FCD \downarrow	AtomStable	MolStable	Bond length \downarrow	Bond angle \downarrow	Dihedral angle \downarrow			
Unified	0.881	0.993	0.935	7.04E-02	9.84E-03	3.47E-04			
	2.356	0.982	0.872	18.4E-02	123E-03	7.03E-04			

Table 11: Reconstruction accuracy (%) for atom and bond types, and coordinate RMSD on the test sets.

Dataset	Method	Atom Accuracy (%)	Bond Accuracy (%)	Coordinate RMSD (Å)
GEOM-Drugs	GeoLDM	96.9	93.7	0.2526
	UDM-3D, ours	100.0	100.0	0.0008
QM9	GeoLDM	98.6	96.2	0.1830
	ADiT	85.7	-	0.3598
	UDM-3D, ours	100.0	100.0	0.0002

into two independent neural networks: one maintaining the R-Trans architecture and taking 2D information (atom types, bond types) as input, while the other uses a vanilla transformer structure and takes 3D information (atom types and coordinates) as input. Each network generates a separate latent space, with dimensions set to half of the original hyperparameter. The LDM is then trained on the concatenated latent space of these two networks. Table 10 summarizes the results on the QM9 dataset. We observe that UDM-3D with a unified latent space significantly outperforms the separated variant across all 2D and 3D generation metrics. For instance, the FCD score improves from 0.351 to 0.130, and the MolStable metric increases from 0.952 to 0.988. This demonstrates that jointly encoding invariant and equivariant features into a single latent representation enables more effective diffusion modeling, as the LDM can learn a coherent distribution over a unified modality rather than needing to capture complex interactions between two separate modalities. These findings validate our design choice of a unified latent space as crucial for achieving state-of-the-art performance in 3D molecule generation.

Influence of Latent Dimensions. To investigate the impact of latent dimensionality on 3D molecule generation, we conduct an ablation study on the QM9 dataset. Table 7 summarizes the reconstruction errors and the performance metrics across different latent space dimensions (4, 8, 16, and 32). The results indicate that a too-low latent dimension (e.g., 4) fails to capture the necessary information, leading to inferior performance. Increasing the dimension to 16 results in substantial performance improvements, while further increasing to 32 does not yield significant additional gains. This demonstrates that a latent dimension of 16 provides an optimal trade-off between model capacity and reconstruction fidelity.

Influence of R-Trans Layer Depth. To assess the effect of depth in the Relational Transformer encoder in UAE-3D, we conduct an ablation study by varying the number of R-Trans layers (3, 6, 9, and 12) on the QM9 dataset regarding reconstruction error. Table 8 summarizes the results. We observe that reducing the number of transformer layers to 3 leads to a noticeable increase in reconstruction error, while increasing it to 9 or 12 does not yield further improvements, as the reconstruction error is already near-zero. This indicates that a moderate depth of 6 layers is sufficient for capturing the necessary relational information among atoms and bonds for effective reconstruction.

Bonded Distance Loss. Table 9 presents an ablation study on the weight λ for the bonded distance loss on the QM9 dataset. We vary λ to analyze its effect on reconstruction accuracy and the preservation of bond geometries. The results indicate that setting $\lambda = 10$ achieves an optimal balance between compressing coordinate information and maintaining geometric fidelity. This experiment validates the necessity of tuning the bonded distance loss weight to enhance model performance.

Ring Distribution. Following the evaluation protocol of [54], we computed the Percentage (%) of molecular modes in terms of ring distribution for our generated molecules and compared them

Table 12: Comparison of ring distributions between molecules in the QM9 training set and those generated by our model.

	3-Ring	4-Ring	5-Ring	6-Ring	(7+)-Rings
QM9 train set	43.14%	39.20%	39.31%	12.72%	2.92%
Ours	41.69%	38.20%	36.37%	12.90%	3.13%

against the QM9 training set. The results in Table 12 provide a quantitative assessment of how well our model preserves the structural diversity of the training dataset.

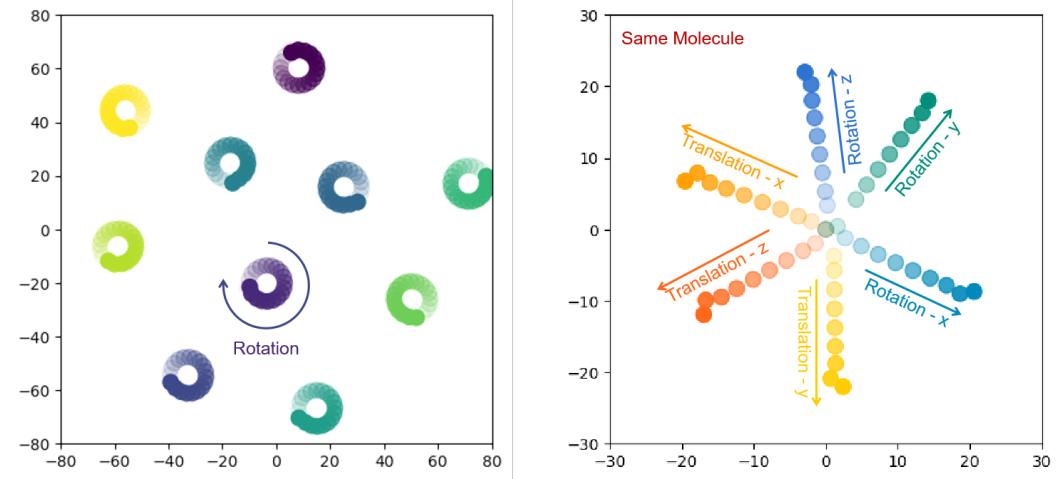


Figure 6: Additional t-SNE visualizations of UAE-3D’s latents under $SE(3)$ augmentations. (a) 360-degree rotation along the z-axis. (b) same molecule under different $SE(3)$ augmentations.

More Visualizations of Latent Space. In addition to Figure 4, we conducted more t-SNE visualizations of UAE-3D’s molecular latent representations to further examine its geometric sensitivity and consistency. In Figure 6(a), we apply 20-step rotations along the same axis (z-axis) to different molecules, completing a full 360-degree rotation. Remarkably, the latent trajectories of all molecule form loops and return to their original positions, reflecting the geometric sensitivity rotational consistency in the learned representations. In Figure 6(b), each color now corresponds to a different $SE(3)$ augmentation applied to the same molecule, rather than to different molecules as in previous plots. We observe clear separation among representations generated by translations and rotations along different axes (x/y/z), indicating that UAE-3D effectively distinguishes between types of geometric transformations in its latent space. These results further validate the structured and disentangled nature of UAE-3D’s latent geometry.

A.2 More Visualizations of Generated Molecules

To further demonstrate the chemical validity and structural diversity of molecules generated by UDM-3D, we provide extended visual comparisons across the GEOM-Drugs and QM9 datasets in Figure 7 and Figure 8.

B More Related Works

Molecular Variational Autoencoders. Variational Autoencoders (VAEs) have been widely explored for generative modeling [16, 55–57], providing a framework to encode data into a structured latent space while enabling both generation and reconstruction [16]. In molecular generation, VAEs have been applied to 2D molecular graphs, as seen in JT-VAE [1], MGCVAE [58], SSVAE [59], and CGVAE [60]. However, these methods have several limitations. First, they primarily focus on 2D

molecular representations and often rely on non-transformer architectures, limiting their scalability for more complex molecular generation tasks [61]. Additionally, VAE-based methods typically assume a simple Gaussian prior, which may fail to accurately model the complex posterior distributions required for effective molecular generation [62, 63]. Notably, masked graph modeling [64] also builds molecular a graph auto-encoder, but aiming for representation learning, instead of molecule generation. With recent advancements in Latent Diffusion Models (LDMs) demonstrating success across multiple domains [65, 66], LDMs have been introduced for 3D molecule and protein generation, as seen in GeoLDM[17] and PepGLAD[18]. These approaches leverage diffusion modeling in latent space, offering improved generative capacity and efficiency.

Comparison to Other Unified Generative Molecule Models. Concurrent with our work, ADiT [38] from FAIR also explores latent diffusion for 3D molecules using a variational autoencoder to compress structures into a unified latent space. However, our approach differs in three key ways: (1) **ADiT does not achieve near-lossless compression.** In contrast, we explicitly evaluate reconstruction quality and show significantly lower errors than ADiT (*cf.* Table 11), which translates to better 3D generation performance (*cf.* Table 2). (2) **ADiT omits chemical bonds as part of the inputs and reconstruction targets, leading to information loss.** Although bond types can be inferred from 3D coordinates, such predictions are not 100% accurate. For example, the QM9 dataset [24] reports 3% inconsistency across inference methods. To address this, we explicitly model bonds using a relational transformer [21] and include bond reconstruction as an objective, offering a more robust and chemically faithful modeling approach. (3) **ADiT provides insufficient evaluation for small molecule generation.** ADiT reports only two metrics (validity and uniqueness) for the QM9 dataset, while our evaluation includes 15 metrics, covering 2D structure, 3D geometry, stability, validity, uniqueness, and other critical aspects (*cf.* Table 1). We conjecture that this is because ADiT is more focused on crystal material generation, instead of small molecule generation.

MolCRAFT [54] also aims to unify multiple molecular modalities for generation. However, it processes atom types and coordinates separately concatenating them while use special design to preserve SE(3) equivariance for coordinates. In contrast, our model encodes atom types, coordinates, and bond types into a single unified latent representation, enabling effortless multi-modal modeling using a uni-modal latent diffusion model – Diffusion Transformer [23]. Moreover, MolCRAFT focuses on the task of structure-based drug design, which is different than our major benchmarks.

Comparison to Other Equivariant 3D Molecule Generation Methods. We adopt SE(3) data augmentations to train a variational autoencoder without explicitly incorporating equivariant architectures, allowing the model to learn geometric symmetries directly from data. This strategy is motivated by AlphaFold3 [28], which demonstrated that SE(3) augmentations are sufficient for general-purpose networks to learn equivariance and invariance. As shown in Table 6, our model achieves near-zero reconstruction errors on 3D molecules subjected to random rotations and translations, showing successful learning of SE(3) equivariance. Compared to models with built-in SE(3) equivariance, such as EDM [19] and JODO [6], our method avoids the architectural complexities of handling multi-modality and enforcing SE(3) equivariant constraints. This design choice allows us to use the highly optimized and parallelizable diffusion model, Diffusion Transformer [23], for latent diffusion modeling. Consequently, we achieve significant efficiency gains, with up to 8.6x faster training and 9.8x faster inference (*cf.* Figure 2(c;d)).

C More Experimental Settings

C.1 Experimental Settings for *De Novo* 3D Molecule Generation

Setup. For the GEOM-Drugs dataset, we use only the ground-state conformer with the lowest energy for each molecule, following the setup in JODO [6]. We adopt the same train/validation/test split ratio of 8:1:1, ensuring a fair comparison. For the QM9 dataset, we use the 100K/10K/10K train/validation/test split. For both datasets, all atoms are shifted to zero center-of-mass before input. All baseline results are either directly borrowed from JODO [6] or reproduced using their released codebase under identical settings.

Hyperparameters. 13 shows the key hyperparameters used for training the UAE-3D and UDM-3D models.

Table 13: Hyperparameters of the UAE-3D and UDM-3D models.

Parameter	UAE	UDM
epochs	2000	10000
atom loss weight (γ_{atom})	1.0	-
bond loss weight (γ_{bond})	1.0	-
coordinate loss weight ($\gamma_{\text{coordinate}}$)	1.0	-
distance loss weight (γ_{distance})	1.0	-
bonded distance loss weight (λ)	10.0	-
KLD loss weight	1e-8	-
batch size	512	
optimizer	AdamW	
learning rate	1e-4	
weight decay	1e-5	
translation augmentation scale	0.1	
encoder hidden size	64	
encoder #heads	8	
encoder #blocks	6	
latent dimension	16	
decoder hidden size	64	
decoder #heads	8	
decoder #blocks	4	
diffusion hidden size	-	512
diffusion #heads	-	8
diffusion #layers	-	8
diffusion mlp ratio	-	4.0
condition drop (p_{drop})	-	0.1

Evaluation Metrics. Our evaluation framework for *de novo* 3D molecule generation encompasses complementary metrics to assess molecular validity, diversity, and geometric fidelity. These metrics operate at two levels:

2D Structural Analysis:

- *Validity & Stability*: Measures adherence to chemical rules:
 - Atom Stability: Percentage of atoms with chemically valid valency, determined by cross-referencing the bond counts with allowed configurations for each atom type (e.g., carbon with 4 bonds). This is evaluated using RDKit’s atom valency parser. Formal charges are taken into account in the valency specification.
 - Molecule Stability: Percentage of molecules with valid valency for all atoms pass the above valency check.
 - Validity & Completeness (V&C): Fraction of fully connected, syntactically correct (SMILES-parsable) molecules, excluding fragmented or hypervalent structures.
- *Diversity*: Quantifies exploration of chemical space:
 - Validity & Uniqueness (V&U): Percentage of unique molecules among valid ones, calculated as the ratio of unique SMILES strings to the total number of valid molecules.
 - Validity & Uniqueness & Novelty (V&U&N): Percentage of unique molecules among valid ones that are also novel, calculated as the ratio of unique SMILES strings to the total number of valid molecules, excluding those present in the training set.
- *Distribution Alignment*: Convert all the valid molecules into SMILES strings and compare generated/test distributions on the MOSES benchmark [3]:
 - Fréchet ChemNet Distance (FCD): Similarity between generated and reference distributions using activations from the penultimate layer of ChemNet [67], a pretrained molecular property predictor. Lower scores indicate better alignment.
 - Similarity to the nearest neighbor (SNN): Tanimoto similarity between the generated molecules and their nearest neighbors in the test set, calculated using Morgan fingerprints.
 - Fragment Similarity (Frag): Tanimoto similarity between the generated molecules and their nearest neighbors in the test set, calculated using BRICS [68] fragments.

- Scaffold Similarity (Scaf): Frequencies of Bemis-Murcko scaffolds [69] in the generated molecules compared to the test set.

3D Geometric Analysis:

- *Conformer Quality*: The 3D coordinates of generated molecules are converted into bond graphs using a predefined lookup of atomic distance thresholds (per atom pair). Based on these bonds, 2D molecular graphs are reconstructed and then evaluated for atom and molecule stability using the same criteria as in 2D Structural Analysis.
 - Atom Stability: Consistency of valency in 3D-derived 2D graphs. Reconstructs 2D bond orders from 3D atomic distances (via tabulated bond-length thresholds) and checks valency compliance.
 - Molecule Stability: Percentage of molecules with valid valency for all atoms pass the above valency check.
 - FCD_{3D} : Same computation as standard FCD, but applied to the reconstructed molecules from 3D coordinates. Captures distributional similarity under geometric constraints.
- *Geometric Fidelity*: Measures spatial arrangement accuracy via Maximum Mean Discrepancy (MMD) for:
 - Bond lengths: MMD between distributions of bond lengths for eight frequent bond types (e.g., C-C, C-N, C-O). Computed using Gaussian kernel with bandwidth tuned via the median heuristic.
 - Bond angles: MMD between the bond angle distributions of generated vs. reference molecules. Triplets of bonded atoms (e.g., C-C-C) define each angle.
 - Dihedral angles: MMD between distributions of dihedral (torsional) angles over four-atom sequences (e.g., C-C-C-C). Captures conformational realism.

Comparison to Other Evaluation Protocols. Discrepancies in evaluation methodologies can significantly impact reported performance metrics. For instance, ADiT’s validity scores are obtained by reconstructing molecules from atomic coordinates using `pymatgen.core.Molecule`, followed by bond inference via RDKit (likely using the `xyz2mol` heuristic [70]). This pipeline ensures topological consistency (i.e., SMILES validity) but does not explicitly enforce chemical stability constraints such as correct valency or bond geometry. In contrast, our evaluation employs RDKit’s `Chem.Mol` for molecule reconstruction, which enforces stricter valency checks and chemical correctness. And our broad evaluation criteria comprehensively measure the validity, diversity, and geometric fidelity of 3D molecule generation.

C.2 Experimental Setting for Conditional 3D Molecule Generation

Setup. We conduct conditional molecule generation on the QM9 dataset following the protocol in EDM [5] and EEGSDE [8]. The properties are evaluated by the QM9 property predictor [24] pre-trained on half of the QM9 dataset. We use the property classifier pre-trained by EDM [5].

Specially, we optimize molecules toward six key electronic properties from quantum chemistry for conditional generation:

- *Electronic Response*:
 - Dipole Moment (μ): Molecular polarity measure
 - Polarizability (α): Induced dipole response to electric fields
- *Thermodynamic Properties*:
 - Heat Capacity (C_v): Energy absorption at constant volume
- *Electronic Structure*:
 - HOMO ($\varepsilon_{\text{HOMO}}$)/LUMO ($\varepsilon_{\text{LUMO}}$): Frontier orbital energies
 - HOMO-LUMO Gap ($\Delta\varepsilon$): Critical for reactivity and conductivity

The evaluation protocol follows a rigorous split-and-validate strategy [5]:

- QM9 training data divided into disjoint subsets D_a (50k) and D_b (50k)
- Property predictor ϕ_c trained exclusively on D_a

- Generated molecules compared against ϕ_c 's predictions on D_b (lower-bound baseline) and random predictions without any relation between molecule and property (upper-bound baseline)
- Performance quantified via Mean Absolute Error (MAE) across all properties

This approach ensures fair assessment of conditional generation without data leakage, with ϕ_c 's D_b performance establishing the theoretical minimum achievable error.

Algorithm 1 Training Algorithm for UAE-3D and UDM-3D

Require: Molecular dataset \mathcal{D} , encoder \mathcal{E} , decoder \mathcal{D} , diffusion model ϵ_θ

- 1: Initialize encoder \mathcal{E} , decoder \mathcal{D} , and diffusion model ϵ_θ
- 2: **Stage 1: Train UAE-3D**
- 3: **while** not converged **do**
- 4: Sample a batch of 3D molecules $\mathbf{G} \sim \mathcal{D}$
- 5: $\mathbf{Z} = \mathcal{E}(\mathbf{G})$ {Encode}
- 6: $\hat{\mathbf{G}} = \mathcal{D}(\mathbf{Z})$ {Decode}
- 7: $\mathcal{L}_{\text{recon}}$ (Eq. 12) {Reconstruction loss}
- 8: $\mathcal{L}_{\text{UAE-3D}}$ (Eq. 13) {VAE loss}
- 9: Update \mathcal{E} and \mathcal{D} using $\mathcal{L}_{\text{UAE-3D}}$
- 10: **end while**
- 11: **Stage 2: Train UDM-3D**
- 12: **while** not converged **do**
- 13: Sample a batch of latent sequences $\mathbf{Z}^{(0)} \sim \mathcal{E}(\mathcal{D})$
- 14: Sample diffusion timestep $t \sim \text{Uniform}(0, 1)$
- 15: $\mathbf{Z}^{(t)} = \sqrt{\bar{\alpha}^{(t)}} \mathbf{Z}^{(0)} + \sqrt{1 - \bar{\alpha}^{(t)}} \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ {Diffusion}
- 16: $\hat{\epsilon} = \epsilon_\theta(\mathbf{Z}^{(t)}, t)$
- 17: $\mathcal{L}_{\text{diffusion}} = \|\hat{\epsilon} - \epsilon\|^2$ {Noise prediction loss}
- 18: Update ϵ_θ using $\mathcal{L}_{\text{diffusion}}$
- 19: **end while**

Algorithm 2 UDM-3D Sampling with CFG

Require: Trained UAE-3D encoder \mathcal{E} , DiT ϵ_θ

Require: Guidance strength w , timesteps $\{t_k\}_{k=1}^K$

Require: Target property c (optional)

- 1: $\mathbf{Z}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ {Initial noise}
- 2: **for** $k = K$ **downto** 1 **do**
- 3: $t \leftarrow t_k$
- 4: **Compute Guidance:**
- 5: $\epsilon_{\text{cond}} \leftarrow \epsilon_\theta(\mathbf{Z}^{(t)}, t, c)$
- 6: $\epsilon_{\text{uncond}} \leftarrow \epsilon_\theta(\mathbf{Z}^{(t)}, t, \emptyset)$
- 7: $\tilde{\epsilon} \leftarrow (1 + w)\epsilon_{\text{cond}} - w\epsilon_{\text{uncond}}$ {CFG blending}
- 8: **Denoise:**
- 9: $\mathbf{Z}^{(t-1)} \leftarrow \text{DDPM_Step}(\mathbf{Z}^{(t)}, \tilde{\epsilon}, t)$ [15]
- 10: **end for**
- 11: **Decode:**
- 12: $\hat{\mathbf{G}} = \langle \hat{\mathbf{F}}, \hat{\mathbf{E}}, \hat{\mathbf{X}} \rangle \leftarrow \mathcal{D}(\mathbf{Z}^{(0)})$ {UAE-3D decoder}

C.3 Pseudo Code

We provide the training and sampling algorithms for UDM-3D in Algorithms 1 and 2, respectively. The training algorithm for UDM-3D involves two stages: training the VAE (UAE-3D) and training the DiT (UDM-3D). The sampling algorithm for UDM-3D with CFG involves iteratively denoising the latent sequence \mathbf{Z} using the DiT and the CFG guidance.

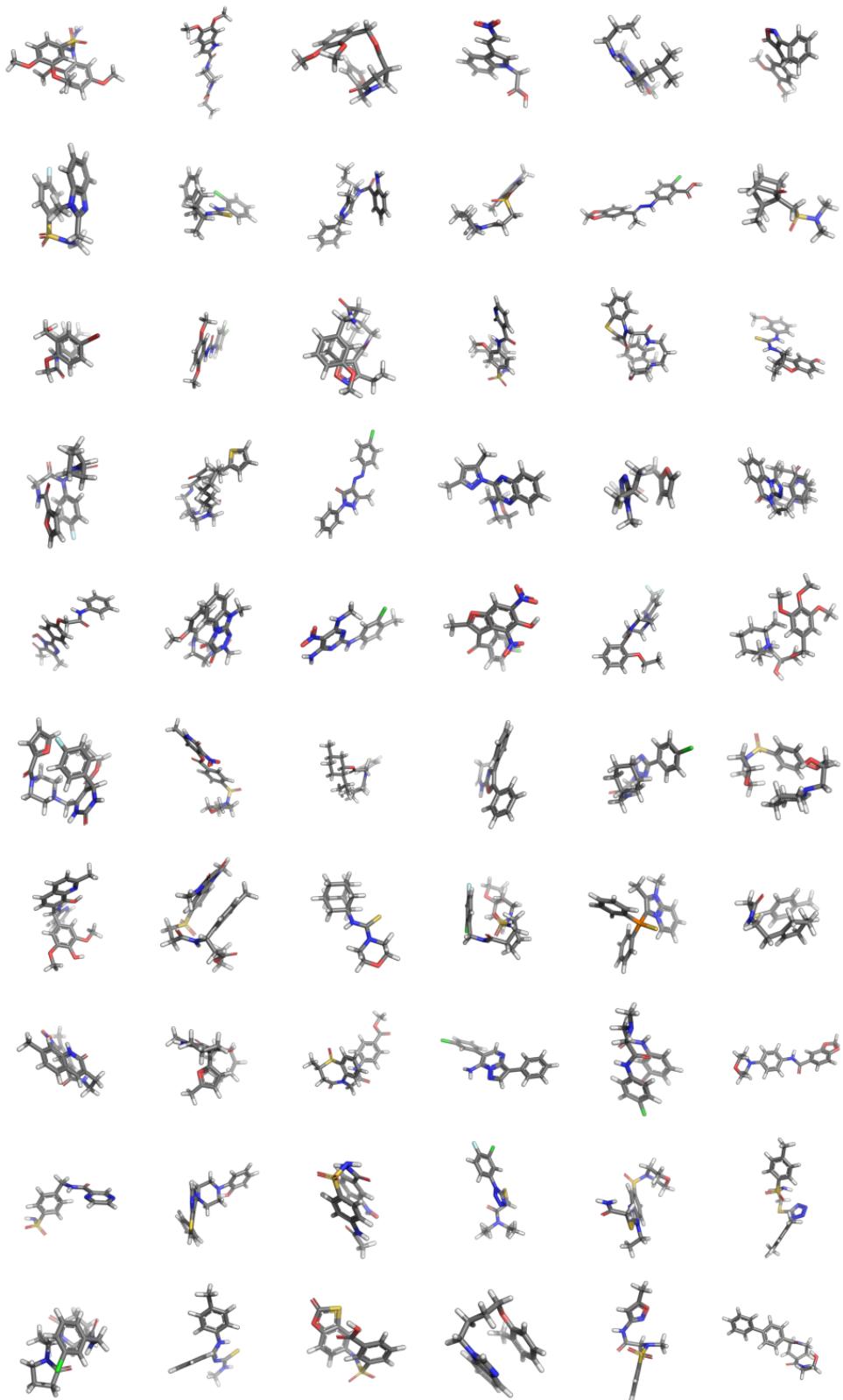


Figure 7: Visualization of random samples generated by UDM-3D on GEOM-Drugs.

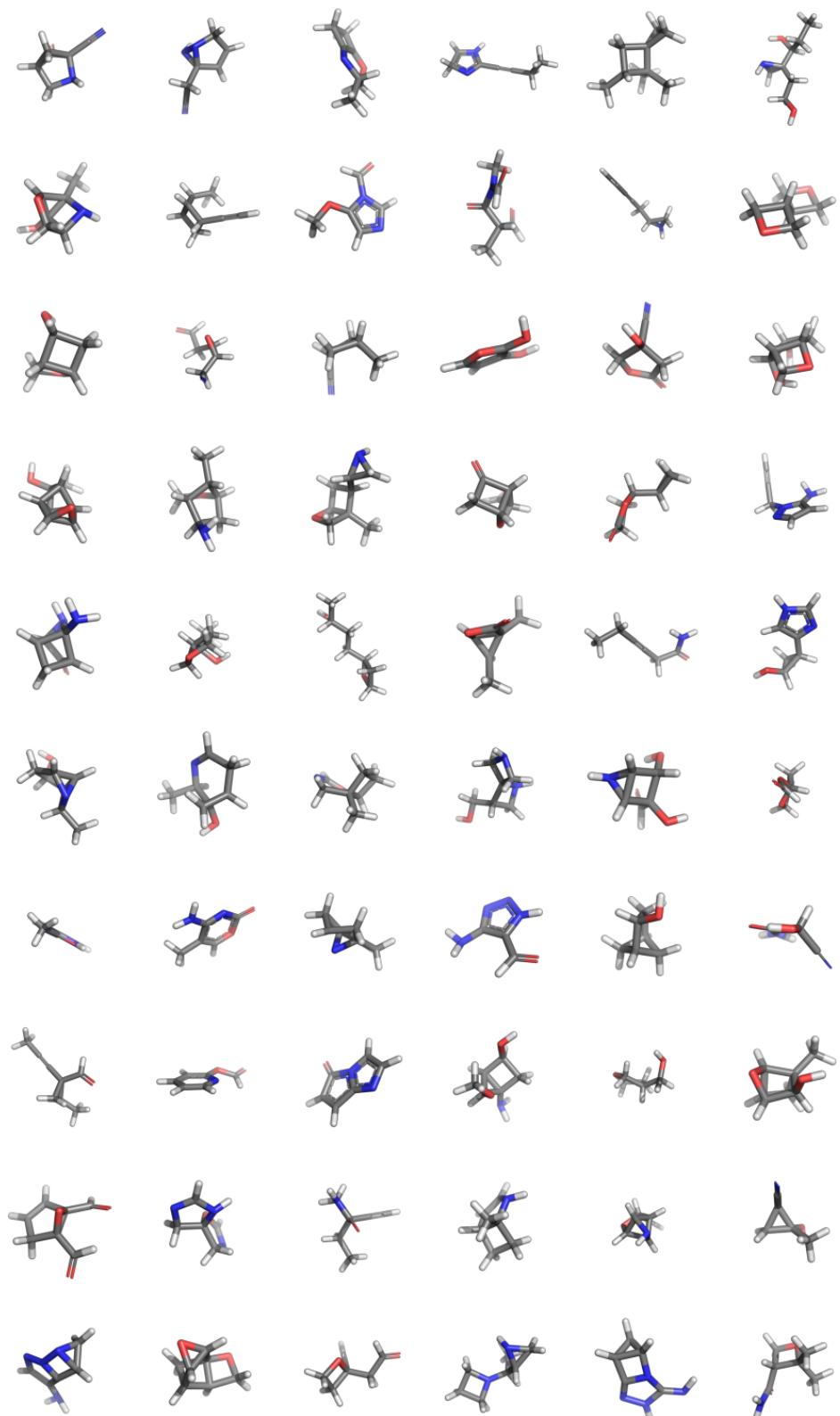


Figure 8: More visualization of random samples generated by UDM-3D on QM9.