

Learning Disentangled Equivariant Representation for Explicitly Controllable 3D Molecule Generation

Haoran Liu^{1,*}, Youzhi Luo^{1,*}, Tianxiao Li², James Caverlee¹, Martin Renqiang Min²

¹Texas A&M University, College Station, TX 77840; ²NEC Laboratories America, Princeton, NJ 08540
{liuhr99, yzluo, caverlee}@tamu.edu, {tili, renqiang}@nec-labs.com

Abstract

We consider the conditional generation of 3D drug-like molecules with *explicit control* over molecular properties such as drug-like properties (e.g., Quantitative Estimate of Druglikeness or Synthetic Accessibility score) and effectively binding to specific protein sites. To tackle this problem, we propose an E(3)-equivariant Wasserstein autoencoder and factorize the latent space of our generative model into two disentangled aspects: molecular properties and the remaining structural context of 3D molecules. Our model ensures explicit control over these molecular attributes while maintaining equivariance of coordinate representation and invariance of data likelihood. Furthermore, we introduce a novel alignment-based coordinate loss to adapt equivariant networks for auto-regressive de-novo 3D molecule generation from scratch. Extensive experiments validate our model’s effectiveness on property-guided and context-guided molecule generation, both for de-novo 3D molecule design and structure-based drug discovery against protein targets.

1 Introduction

Although a lot of progress has been made in AI-powered 2D molecule design (Bengio et al. 2021; Jin, Barzilay, and Jaakkola 2018; Shi et al. 2019; Xie et al. 2021), generating desirable 3D molecules with specific attributes, e.g., drug-like properties (e.g., QED and SA Score) and the ability to bind with targeted proteins, is still challenging and crucial in drug discovery. Given the vastness of the chemical space, estimated to be in the order of 10^{33} (Polishchuk, G, and Varnek 2013), conducting an exhaustive search is impractical. Therefore, our goal is to develop computational methods for direct generation of novel and valid drug-like molecules *conditional* on desired properties.

To perform *conditional* 3D molecule generation, existing auto-regressive models such as G-SchNet (Gebauer, Gastegger, and Schütt 2019) and G-SphereNet (Luo and Ji 2022) can be fine-tuned using subsets of favorable data, selected based on specific threshold criteria. However, these models are limited by their inability to generate molecules

with desired property values such as asphericity, QED, SA Score, and binding to target proteins. On the other hand, equivariant diffusion models (Hoogetboom et al. 2022; Qiang et al. 2023; Xu et al. 2023) address this limitation by directly incorporating property values as additional inputs. Nevertheless, they encounter a different challenge due to lacking an explicit latent space for direct manipulation of latent variables, which is required for assigning or anchoring distinct properties to the noise/latent space. Therefore, they are unable to explicitly control molecule attributes in the generation process. However, explicit control over 3D molecule generation is crucial in real-world drug design scenarios. Specialists often need to optimize certain properties while maintaining others. For instance, improving a drug’s synthetic accessibility without altering its binding affinity to a target. This precision demands a generation process with a manipulable latent space, enabling targeted property modifications.

To achieve *explicit control* over 3D molecule generation, we propose to incorporate disentangled representation learning (Bengio, Courville, and Vincent 2013; Higgins et al. 2016; Kim and Mnih 2018) into this task. Specifically, we factorize the latent space of our generative model into two disentangled aspects: one representing the desired molecular properties, and the other capturing the remaining structural context related to 3D molecular validity. Notably, almost all existing methods only focus on the property-guided side of 3D molecule generation, neglecting the control over the remaining component of the latent space, i.e., the complementary factors to molecular properties. In contrast, our model supports two distinct generation modes: **property-targeting generation**, for which we generate 3D molecules with specific, desirable properties for property-targeting 3D generation, and **context-preserving generation**, for which we generate 3D molecules allowing optimization of targeted molecular properties while maintaining the molecule’s fundamental architecture. These two capabilities enabled by our model are essential for many real-world drug design scenarios, where minor modifications can critically impact a drug’s efficacy and safety. Especially, the *unique capability for context-preserving generation* distinguishes our model from all existing 3D molecule generation methods.

To realize these capabilities, we propose an *E(3)-equivariant Wasserstein autoencoder* (E3WAE) model ar-

*Most of this work was done when Haoran and Youzhi were interns at NEC Labs America.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

architecture based on the above disentangled representation learning framework for controllable 3D molecule generation in this paper. The disentanglement is achieved by two components: a Wasserstein regularization loss (Tolstikhin et al. 2018) to force the independence between property latent variables and context latent variables, and a prediction head to ensure that the property latent variables contain property-relevant information. Specifically, our model adopts a fragment-based auto-regressive approach, tailored for generating large-scale, drug-like molecules. However, it is infeasible to directly apply equivariant networks (Batatia et al. 2022; Geiger and Smidt 2022; Satorras, Hoogeboom, and Welling 2021) to auto-regressive 3D molecule generation without any external reference structures. To overcome this challenge, we devise a novel coordinate loss function with structure alignment to guarantee equivariance of coordinate representation and invariance of data likelihood. Leveraging E(3)-equivariant networks, disentangled representation learning, and the coordinate loss, E3WAE outperforms previous methods on property-targeting generation and performs effective context-preserving generation on which previous methods fail.

To summarize, our contributions are as follows: 1) We propose the *first* disentangled representation learning framework for both *de-novo* and **structure-based** 3D molecule generation, achieving *explicit control* over molecule property and structure context. This benefits real-world drug design scenarios where it is often necessary to fine-tune a specific property while keeping others unchanged. 2) We design a novel auto-regressive E(3)-equivariant Wasserstein autoencoder with a new coordinate prediction loss to adapt equivariant networks for *de-novo* 3D molecule generation. 3) We validate our model’s effectiveness in property-targeting and the novel context-preserving generation tasks for both *de-novo* 3D molecule generation and structure-based drug design.

2 Related Works

De-novo Molecular Generation in 3D. Recently, there has been considerable progress in the field of 3D molecule generation using deep learning techniques. Various approaches have been explored to accurately represent and generate 3D molecular structures. Some have focused on learning joint distributions of geometric features (Garcia Satorras et al. 2021; Luo and Ji 2022). Others, such as G-SchNet (Gebauer, Gastegger, and Schütt 2019), have employed auto-regressive models with symmetry constraints to sample 3D molecules directly in their spatial configurations. A notable advancement is the use of equivariant diffusion models (Hoogeboom et al. 2022; Qiang et al. 2023; Xu et al. 2023), which have shown significant promise in generating high-quality molecules. These techniques have been further applied in specific generation tasks with external conditioning, such as reference ligands (Adams and Coley 2023; Liu et al. 2022; Powers et al. 2022; Guan et al. 2023) and linkers (Huang et al. 2022; Igashov et al. 2022; Imrie et al. 2020). Despite these advancements, a common limitation in most existing models is the lack of *explicit control* over the attributes of the generated molecules. This poses a challenge for appli-

cations where there is a *need to modify certain attributes of the molecules*, such as a specific chemical property, *while preserving other aspects* like overall shape and composition. Our research aims to address this gap by *explicitly controlling* the generation of molecules, where higher-level semantics can be precisely manipulated to meet specific criteria.

Disentangled Representation Learning. The design of disentangled representation learning (DRL) tasks varies based on the nature of data involved. For instance, DRL tasks involve decoupling “style” and “content” in text (Cheng et al. 2020) or images (Kotovenko et al. 2019), or distinguishing between “static” and “dynamic” components in videos (Zhu et al. 2020; Han et al. 2021). Such disentangled representations are pivotal in applications like style transfer (Cheng et al. 2020; Kotovenko et al. 2019; Lee et al. 2018) and conditional generation (Denton et al. 2017; Zhu et al. 2018). To achieve disentanglement, many regularization techniques are developed. These techniques include managing the capacity of latent space (Burgess et al. 2018; Higgins et al. 2016), implementing adversarial loss (Deng et al. 2020), employing cyclic reconstruction (Lee et al. 2018), imposing mutual information constraints (Chen et al. 2016; Cheng et al. 2020; Zhu et al. 2020), and integrating self-supervising auxiliary tasks (Zhu et al. 2020). In molecular science, DRL has been applied to molecule generation using SMILES sequences (Mollaysa, Paige, and Kalousis 2020) and 2D molecular graphs (Du et al. 2022). However, to the best of our knowledge, its application in *de-novo* 3D molecule generation and structure-based drug design has not been explored.

3 Preliminaries

Problem Formulation. To generate *large-scale drug-like* molecules, our method adopts an auto-regressive, fragment-based 3D generation approach. Let \mathcal{G} be the space of 3D molecular graphs, where each 3D molecular graph $G \in \mathcal{G}$ consists of the fragment node set \mathcal{V} , the edge set \mathcal{E} , and the fragment coordinate matrix \mathcal{R} . Specifically, each fragment represents a combination of several atoms and bonds, e.g., a benzene ring could be a fragment that includes six carbon atoms and aromatic bonds. Correspondingly, each fragment $i \in \mathcal{V}$ is associated with a node feature v_i and a position vector \mathbf{r}_i , which corresponds to the i -th line of \mathcal{R} and represents the center coordinates of the fragment. Following fragment-based generation (Bengio et al. 2021; Jin, Barzilay, and Jaakkola 2018; Qiang et al. 2023), we use edge $e_{ij} \in \mathcal{E}$ to indicate that two fragments $i, j \in \mathcal{V}$ share a bond/atom. This fragmentization inherently includes assembled rules for neighboring fragments and enables us to model a substantial portion of the chemical space with a reasonably sized fragment vocabulary. The generation process is iterative; at each step t , the next graph state is predicted as $\mathcal{G}_{t+1} = \mathcal{M}(\mathcal{G}_t)$, where \mathcal{M} is the auto-regressive generation model. The fragmentization details are in Appendix 7. Building upon this, our model is designed for explicit control over molecular attributes while ensuring both E(3)-invariance and equivariance properties.

Explicit Control. We define a generative model with explicit control as $\mathcal{M} : \mathcal{Y}_1 \times \mathcal{Y}_2 \rightarrow \mathcal{X}$, where \mathcal{Y}_1 and \mathcal{Y}_2 rep-

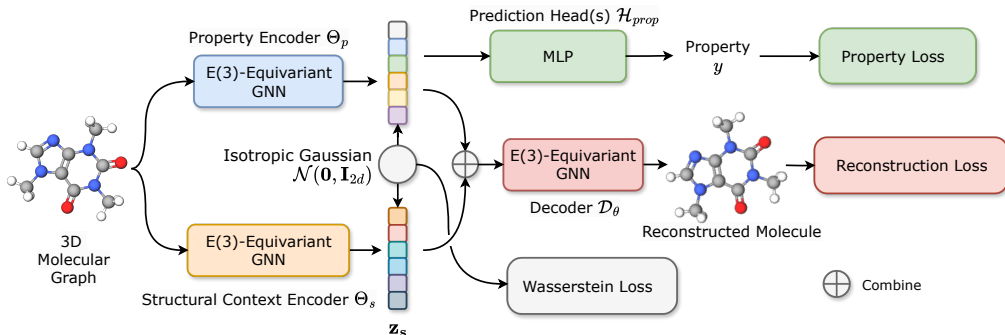


Figure 1: An illustration of the proposed E3WAE framework. A 3D molecule is encoded into two disentangled latent variables: the property variable z_p and the structural context variable z_s with two E(3)-equivariant encoders, respectively. The latent variables are then combined to reconstruct the molecule in an auto-regressive manner. A prediction head with the supervision of property labels is attached to z_p to ensure z_p carries information related to the property. Disentanglement of the property and structure variables is achieved through a Wasserstein autoencoder regularization loss and minimization of the MMD distance against an isotropic Gaussian distribution. The overall training objective combines a reconstruction loss, a property prediction loss, and the Wasserstein loss.

represent the attribute spaces of two different attributes, and \mathcal{X} is the space of generated objects. The model \mathcal{M} is considered to have explicit control if, for any given attributes $y_1 \in \mathcal{Y}_1$ and $y_2 \in \mathcal{Y}_2$, it consistently generates an object $x \in \mathcal{X}$ reflecting these attributes. In our model, \mathcal{Y}_1 is the space of target molecular properties, and \mathcal{Y}_2 refers to the structural context of molecules. Manipulating y_1 allows for altering molecular properties and maintaining their structure context, whereas adjusting y_2 enables refining molecule structures and maintaining their chemical properties.

E(3)-Invariance and Equivariance. E(3) or 3D Euclidean group is the group generated by all 3D rotations, translations and reflections in the 3D space. In the context of *de-novo* 3D molecule generation from scratch, i.e. without external reference structure, the goal is to design models maintaining equivariance of coordinate representation and invariance of data likelihood. First, the coordinates of a molecule’s nodes are *equivariant* to the positions of other nodes in the 3D molecule. That means, when generating the coordinates r_t for an atom at iteration t , if we rotate or translate the current input structure \mathcal{R}_t , then r_i needs to be rotated or translated correspondingly. Formally, consider a rotation matrix $\mathbf{R} \in SO(3)$ and a translation vector $\tau \in \mathbb{R}^3$, for the coordinate generation model \mathcal{M}^r , we have $\mathbf{R}r_t + \tau = \mathcal{M}^r(\mathbf{R}\mathcal{R}_t + \tau)$. In contrast, the likelihood $\pi(r_i|\mathcal{R}_t)$ should be *invariant* to rotations and translations as they do not change the 3D structure, so that $\pi(\mathbf{R}r_t + \tau|\mathbf{R}\mathcal{R}_t + \tau) = \pi(r_t|\mathcal{R}_t)$, for any rotation $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ or translation $\tau \in \mathbb{R}^3$.

4 Methodology

We propose an *E(3)-equivariant Wasserstein autoencoder* (E3WAE) model, where the generation of the 3D molecule is factorized into two disentangled factors: the **property** and the **structural context**. The former variable comprises the chemical property of the 3D molecule while the latter refers to all other 3D structure patterns that do not relate to the property yet reflect the chemical constraints within the molecular chemical space. This factorization allows for *explicit control* during the generation process. In addition, we introduce a

novel coordinate prediction loss as part of the reconstruction loss to enable auto-regressive 3D molecule generation with equivariant networks. An overview of our model framework is shown in Figure 1.

4.1 E(3)-Equivariant Disentangled Encoder

We use two separate E(3)-equivariant GNN encoders to extract invariant and equivariant latent variables, respectively:

$$\mathbf{z}_{h,e}, \mathbf{z}_{v,e} = \Theta_e(\mathcal{G}), \quad (1)$$

where $e = \{p, s\}$ represents “property” or “structural context”. The invariant latent variable associate with node i , $z_{h,e}^i \in \mathbf{z}_{h,e}$, has a dimensionality of d_h . Correspondingly, the equivariant latent variable $\mathbf{z}_{v,e}^i \in \mathbf{z}_{v,e}$ is dimensioned in $\mathbb{R}^{d_v \times 3}$.

In this paper, we opt for E(3)-equivariant Vector Neurons Multilayer Perceptron (VN-MLP) (Deng et al. 2021) and Mixed-Features Message Passing (MF-MP) (Huang et al. 2022) as building blocks for both branches of encoders Θ_e . These choices are due to their proven effectiveness in integrating both invariant and equivariant features, ensuring the property space’s invariance and the coordinate space’s equivariance. However, it is important to note that our model is compatible with any E(3)-equivariant architecture.

For the *property* branch, the property latent variables $\mathbf{z}_{h,p}$ and $\mathbf{z}_{v,p}$ are first derived by the property encoder $\Theta_p(\mathcal{G})$. Subsequently, a Readout function is used to obtain graph-level representations. The Readout function is implemented as either the average or the summation of all node embeddings. In order to ensure that these latent variables carry information related to the property of 3D molecules, we introduce an auxiliary prediction head \mathcal{H}_{prop} that takes the aggregated $\mathbf{z}_{h,p}$ as input and predicts the target property value of the 3D molecule:

$$\hat{y} = \mathcal{H}_{prop}(\text{Readout}(\mathbf{z}_{h,p})), \quad (2)$$

For the *structure context* branch, the context latent variables $\mathbf{z}_{h,s}$ and $\mathbf{z}_{v,s}$ are encoded using the structural encoder

$\Theta_s(G)$. To ensure that these variables capture comprehensive information about the molecule, specifically excluding patterns related to the target property, we employ an auto-regressive molecule reconstruction loss. Details of this process are further discussed in Section 4.3.

4.2 Disentanglement of the Latent Space

To achieve the disentanglement between property and context latent variables, we introduce a Wasserstein autoencoder regularization loss following Tolstikhin et al. (2018). This approach involves minimizing the Maximum Mean Discrepancy (MMD) between the distribution of latent variables and an isotropic multivariate Gaussian prior, denoted as $\mathbf{z} \sim P_{\mathbf{z}}$. Specifically, for the invariant latent variables $\mathbf{z}_h \sim Q_{\mathbf{z}}^h$, where $\mathbf{z}_h = \text{concat}(\mathbf{z}_{h,p}, \mathbf{z}_{h,s})$, and an isotropic Gaussian prior $P_{\mathbf{z}}^h = \mathcal{N}(\mathbf{0}, \mathbf{I}_{2d_h})$, we compute the disentanglement loss for invariant variables as

$$\mathcal{L}_{\text{Dis}}^h = \text{MMD}(P_{\mathbf{z}}^h, Q_{\mathbf{z}}^h). \quad (3)$$

For the equivariant latent variables, $\mathbf{z}_v = \text{concat}(\mathbf{z}_{v,p}, \mathbf{z}_{v,s})$ and $\mathbf{z}_v \in \mathbb{R}^{2d_v \times 3}$, we maintain independence along the $2d_v$ axis while allowing for covariance along the remaining dimension. To this end, we sample *three* isotropic Gaussian priors from $P_{\mathbf{z}}^v = \mathcal{N}(\mathbf{0}, \mathbf{I}_{2d_v})$ and calculate the corresponding disentanglement Wasserstein loss as:

$$\mathcal{L}_{\text{Dis}}^v = \text{MMD}(\mathbf{P}_{\mathbf{z}}^v, \mathbf{Q}_{\mathbf{z}}^v), \quad (4)$$

where $\mathbf{P}_{\mathbf{z}}^v \in \mathbb{R}^{2d_v \times 3}$ is the combined Gaussian priors and $\mathbf{Q}_{\mathbf{z}}^v \in \mathbb{R}^{2d_v \times 3}$ is the distribution of the equivariant latent variables \mathbf{z}_v . The total disentanglement loss is then calculated as the sum of these components: $\mathcal{L}_{\text{Dis}} = \mathcal{L}_{\text{Dis}}^h + \mathcal{L}_{\text{Dis}}^v$.

For MMD estimation, we take an input batch of latent variables $\{\mathbf{z}_i\}_{i=1,\dots,m}$ with batch size m . Corresponding samples $\{\tilde{\mathbf{z}}_i\}_{i=1,\dots,m}$ are randomly drawn from the Gaussian prior, maintaining the same sample size. The MMD is then calculated using the linear time unbiased estimator (Gretton et al. 2012; Sutherland et al. 2017) as

$$\text{MMD}(P_{\mathbf{z}}, Q_{\mathbf{z}}) = \frac{1}{\lfloor m/2 \rfloor} \sum_{i=1}^{\lfloor m/2 \rfloor} [k(\mathbf{z}_{2i-1}, \mathbf{z}_{2i}) + k(\tilde{\mathbf{z}}_{2i-1}, \tilde{\mathbf{z}}_{2i}) - k(\mathbf{z}_{2i-1}, \tilde{\mathbf{z}}_{2i}) - k(\mathbf{z}_{2i}, \tilde{\mathbf{z}}_{2i-1})], \quad (5)$$

where k is the kernel function, implemented using a radial basis function (RBF) kernel with $\sigma = 1$.

Minimizing this loss aligns the joint distribution of the latent embeddings with the isotropic normal distributions, so that the property and context latent variables are independent. Additionally, the Gaussian shape of the latent space facilitates smooth interpolation, effective regularization, and enhanced generation diversity (Kingma and Welling 2014; Tolstikhin et al. 2018). A theoretical analysis supporting the disentanglement guarantee is provided in Appendix 8.

4.3 Decoder and 3D Molecule Graph Reconstruction

The decoder $\mathcal{D}(\mathcal{G}|\mathbf{z}_h, \mathbf{z}_v)$ reconstructs 3D molecule graph \mathcal{G} fragment-by-fragment in an auto-regressive manner. Specif-

ically, we employ another E(3)-equivariant GNN as the decoder. We incorporate the connectivity rules between fragments by masking out invalid edges. For the auto-regressive reconstruction, our approach is inspired by the *focus and expand* molecule generation procedure in previous works (Ceylan and Gutmann 2018; Huang et al. 2022). Notably, our method is significantly different from previous works. We use fragments as nodes and introduce a novel coordinate prediction loss (see Section 4.4) specifically designed to optimize equivariant networks. This approach enables the generation of molecules using equivariant networks without relying on any external conditioning, such as reference structures like linker designs and pocket-based generation. A detailed algorithm for this decoding process is in Appendix 9.

Specifically, the decoding process includes the following steps: (i) **Node type prediction**: Initially, fragment types $\{x_i\}_{i=1}^n$ for all n nodes in \mathcal{G} are determined from latent variables \mathbf{z}_h and \mathbf{z}_v . The fragment type logits are obtained from latent variables through a self-attention mechanism and an MLP, and these logits are then used to sample all fragment types. Once the node types are sampled, their embeddings are concatenated with the corresponding latent variables \mathbf{z}_h^i for subsequent steps. (ii) **Focus queue initialization**: A focus queue Q is then initialized with a randomly chosen node from \mathcal{G} . (iii) **Focus and expand iterations**: For each *focus* node f popped from Q , the decoder predicts an *expand* edge connecting f to another node. If the connected node u is not a stop node or being linked for the first time, its coordinate (x_u, y_u, z_u) is predicted. Node u is then added to Q if unvisited. This process repeats until a stop node is reached, marking f as *visited*. The order of node focus and edge connection is determined using Breadth-First Search, facilitating teacher-forcing training. The reconstruction process ends when Q is empty, ensuring that all nodes in the connected component of \mathcal{G} have been considered for expansion.

Concretely, in each iteration t of the focus and expand phase, we have the currently reconstructed subgraph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t, \mathcal{R}_t)$. Initially, the latent node embeddings are updated as $\hat{\mathbf{z}}_h, \hat{\mathbf{z}}_v$ with a MF-MP layer. Then the edge logits between the focus node f and any node i are obtained with

$$e_{f,i} = \Phi \left(\hat{\mathbf{z}}_h^f, \hat{\mathbf{z}}_h^i, \|\hat{\mathbf{z}}_v^f\|, \|\hat{\mathbf{z}}_v^i\|, \mathbf{m}_{f,i}, \sum_{j \in \mathcal{V}_t} \hat{\mathbf{z}}_h^j, \sum_{j \in \mathcal{V}_t} \hat{\mathbf{z}}_h^j \right), \quad (6)$$

where Φ is a feed forward network and $\mathbf{m}_{f,i} \in \{0, 1\}$ indicates whether fragments f and i can be connected, based on feasible chemical bonding rules. The softmax function is applied to the edge logits to determine the probabilities for each edge. Node u is then determined by identifying which node i has the highest probability.

To predict the coordinate for the newly connected node u , we take the geometric center $\bar{\mathbf{r}}_t$ of current subgraph \mathcal{G}_t as a reference point and predict a displacement of node u related to the reference point. Initially, we compute two sets of pair-wise interactions in the current connected graph as

$$p_{ij}^{(k)} = \Psi(\hat{\mathbf{z}}_h^i, \hat{\mathbf{z}}_h^j, \langle \mathbf{W}^{(k)} \hat{\mathbf{z}}_v^i, \mathbf{W}^{(k)} \hat{\mathbf{z}}_v^j \rangle), \quad k = 1, 2 \quad (7)$$

where Ψ is a feed forward network and $\mathbf{W}^{(k)}$ is a learnable linear transformation. The displacement of node u is then calculated by:

$$\mathbf{d}_u = \sum_{j \in \mathcal{V}_t} p_{uj}^{(1)} (\mathbf{r}_j - \bar{\mathbf{r}}_t) + \Omega_1 \left(\sum_{j \in \mathcal{V}_t} p_{uj}^{(2)} \Omega_2(\hat{\mathbf{z}}_v^u, \hat{\mathbf{z}}_v^j) \right) \quad (8)$$

where Ω_1 and Ω_2 are VN-MLP layers. Finally, the predicted coordinates of node u is obtained by $\hat{\mathbf{r}}_u = \mathbf{d}_u + \bar{\mathbf{r}}_t$.

4.4 Training Objectives

The proposed E3WAE network is trained by minimizing a weighted sum of three individual losses:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Prop}} + \alpha \mathcal{L}_{\text{Dis}} + \beta \mathcal{L}_{\text{Recon}}, \quad (9)$$

where α and β are the trade-off weights for the losses. Specifically, \mathcal{L}_{Dis} is introduced in Section 4.2. In addition, for the auxiliary property prediction head $\mathcal{H}_{\text{prop}}$ attached to property encoder Θ_p in Section 4.1, we adopt the L1 loss as the property prediction loss:

$$\mathcal{L}_{\text{Prop}} = \|y - \hat{y}\|_1, \quad (10)$$

where y denotes the ground truth value of the target property. Finally, for the decoder in Section 4.3, we adopt a reconstruction loss which consists of three parts:

$$\mathcal{L}_{\text{Recon}} = \mathcal{L}_{\text{NodeType}} + \mathcal{L}_{\text{Edge}} + \mathcal{L}_{\text{Coords}}. \quad (11)$$

Initially, a cross-entropy loss $\mathcal{L}_{\text{NodeType}}$ is used for a classification task to accurately determine the types of nodes. Following this, another cross-entropy loss, $\mathcal{L}_{\text{Edge}}$, is applied to predict the edges in each iteration of the process. Finally, we incorporate a log-MSE loss (Yu 2020) as the coordinate prediction loss $\mathcal{L}_{\text{Coords}}^t$ at each iteration t :

$$\mathcal{L}_{\text{Coords}}^t = \log(\sum_{i=1}^m f_i \|\bar{\mathbf{r}}_i - \mathbf{r}_i\|^2) / (\sum_{i=1}^m f_i), \quad (12)$$

where f_i is a binary flag indicating the presence of a newly added node in the i -th subgraph $\mathcal{G}_t^{(i)}$ that is not a stop node.

However, we cannot apply the log-MSE loss directly. Specifically, for the first iteration $t = 1$, there are only two nodes in the current subgraph, in which $|\mathcal{V}_t| = 2$ and $\mathcal{R}_t \in \mathbb{R}^{2 \times 3}$. The only constraint to the current 3D structure is the distance between the two existing nodes. Thus, the ground truth coordinates can be considered as any point on a sphere with distance $d_t = \|\mathbf{r}_u - \bar{\mathbf{r}}_t\|$ to the reference point coordinate $\bar{\mathbf{r}}_t = \mathbf{r}_f$, where \mathbf{r}_u denotes the ground truth coordinate of newly added node u and \mathbf{r}_f is the coordinate of the focus and the only node f in previous subgraph. Moreover, for the situation that there are three nodes in the current subgraph, where $|\mathcal{V}_t| = 3$ and $\mathcal{R}_t \in \mathbb{R}^{3 \times 3}$, the ground truth coordinate is considered equivalent to any point with a distance $d_t = \|\mathbf{r}_u - \bar{\mathbf{r}}_t\|$ to $\bar{\mathbf{r}}_t$ and a angle $\theta_t = \arccos(\mathbf{s}_t, \mathbf{s}_{uf})$. Here, we use $\mathbf{s}_t, \mathbf{s}_{uf}$ to denote the unit vectors of $\mathbf{r}_u - \bar{\mathbf{r}}_t$ and $\mathbf{r}_u - \mathbf{r}_f$, respectively. Therefore, if we use the log-MSE coordinate loss directly, all these situations will be neglected and the symmetric-invariance property of the geometric space will be violated.

To adapt E(3)-equivariant networks to *de-novo* molecule generation task *without external reference structure*, we propose to align the coordinates with Kabsch algorithm (Kabsch 1976) and then calculate the coordinate loss with transformed coordinates for the case when there are less than or equal to three nodes in any samples in this batch. Concretely, we calculate a rotation matrix $\mathbf{R} \in SO(3)$ and a translation vector $\tau \in \mathbb{R}^3$ for the maximum alignment between generated coordinates $\hat{\mathcal{R}}_t = [\hat{\mathbf{r}}_i]_{i \in \mathcal{V}_t}$ and ground truth coordinates \mathcal{R}_t . Then transformed coordinates of generated nodes are obtained by

$$\tilde{\mathbf{r}}_i = \mathbf{R} \hat{\mathbf{r}}_i + \tau, \quad \forall i \in \mathcal{V}_t. \quad (13)$$

Thus, for samples with three or fewer nodes at the current iteration, the coordinate loss is:

$$\mathcal{L}_{\text{Coords}}^{t,i} = \sum_{j \in \mathcal{V}_{t,i}} \|\tilde{\mathbf{r}}_j - \mathbf{r}_j\|^2. \quad (14)$$

4.5 Generation with Explicit Control

In the generation phase, we sample latent variables \mathbf{z}_h and \mathbf{z}_v and set the maximum number of fragments to N . Note that the *exact* number of fragments might be smaller than N . Then the generation process follows the same decoding process introduced in Section 4.3 but differs by operating without teacher forcing, relying solely on the model’s self-guided predictions. Notably, a unique feature of our E3WAE model is the disentangled latent space, which enables *explicit control* over two key aspects of molecule generation: (i) **property-targeting generation**: In this approach, predefined property latent variables $\mathbf{z}_{h,p}, \mathbf{z}_{v,p}$ are combined with sampled or template context latent variables $\mathbf{z}_{h,s}, \mathbf{z}_{v,s}$. This combination is then fed into the decoder, producing new molecules $\hat{\mathcal{G}} = \mathcal{D}(\mathcal{G}|\mathbf{z}_{h,p}, \mathbf{z}_{v,p})$ with targeted properties; (ii) **context-preserving generation**: Conversely, this mode uses predefined context latent variables $\mathbf{z}_{h,s}, \mathbf{z}_{v,s}$ combined with either sampled or template property latent variables $\mathbf{z}_{h,p}, \mathbf{z}_{v,p}$. Using such latent variables, the decoder can generate new molecules $\hat{\mathcal{G}} = \mathcal{D}(\mathcal{G}|\mathbf{z}_{h,s}, \mathbf{z}_{v,s})$ that refine certain properties while maintaining the core molecule framework. The predefined latent variables are obtained either by directly using or by performing interpolation or extrapolation with latent variables from existing molecules. For simplicity, this work adopts the first approach.

Atom Conformation Assembling. Once the 3D molecular graph is complete at the fragment level, including all fragment types, links, and center coordinates, we proceed to determine the atom-level coordinates. Following Qiang et al. (2023), we first randomly pick a fragment and explore all feasible links to its neighbors, selecting the one nearest to our pre-defined fragment center. We then use RDKit (Landrum 2016) to model each potential connection, evaluating their suitability based on root-mean-square deviation (RMSD) from the center. This procedure is iteratively applied to build the structure fragment by fragment. Finally, we align these local structures within the molecular framework using the Kabsch algorithm (Kabsch 1976), adjusting the RDKit coordinates to match target positions.

Table 1: Results on property-targeting generation. We tested the difference using mean squared error (MSE) and mean absolute error (MAE) between the given reference and the actual properties of the generated 3D molecules.

		Asphericity		QED		SAS		logP	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GEOM	EDM	0.626	0.455	0.113	0.285	12.395	3.385	5.704	1.903
	HierDiff	0.176	0.406	0.12	0.289	2.618	1.347	4.124	1.572
	Ours	0.095	0.246	0.072	0.221	1.563	1.002	4.490	1.630
Cross-Docked 2020	EDM	0.109	0.274	0.147	0.309	11.244	3.205	5.715	1.957
	HierDiff	0.107	0.268	0.089	0.278	2.364	1.468	5.752	1.897
	Ours	0.100	0.259	0.062	0.205	2.356	1.243	4.244	1.644

5 Experiments

In our work, we focus on generating large-scale drug-like molecules. We follow the experimental setup in Qiang et al. (2023) and conduct experiments mainly on two benchmark datasets for 3D molecule generation: GEOM-Drugs (Axelrod and Gomez-Bombarelli 2022) and CrossDocked2020 (Francoeur et al. 2020). GEOM-Drugs contains 304k drug-like molecules while CrossDocked2020 consists of about 100k 3D ligand structures, each derived from a protein-ligand complex. Our model is trained using a subset of 50,000 molecular structures from each dataset, specifically selecting those with the lowest energy conformations for each molecule. We compare our model with three well-established baseline methods, including one auto-regressive method, G-SphereNet (Luo and Ji 2022) and two equivariant diffusion-based methods, EDM (Hooeboom et al. 2022) and HierDiff (Qiang et al. 2023). Hyperparameters and other implementation details are provided in Appendix 11.

5.1 Property-targeting 3D Generation

In AI-enhanced drug discovery, a crucial focus is the generation of 3D molecules with specific, desirable properties. This task is commonly referred to as *property-targeting 3D generation*. Following the experiment settings in HierDiff (Qiang et al. 2023), we select asphericity, QED, SAS, and logP as key properties for conducting practical drug discovery-related conditional generation. Asphericity assesses how much a molecule’s shape deviates from a sphere, influencing its biological interactions. Quantitative Estimate of Drug-likeness (QED) gauges a molecule’s similarity to known drugs, indicating its potential efficacy as a drug candidate. The Synthetic Accessibility Score (SAS) evaluates the ease of synthesizing a molecule, a crucial factor in drug development. Lastly, logP measures a compound’s lipophilicity, which is essential for understanding the compound’s absorption and distribution in the body. Each of these properties plays a significant role in determining a molecule’s suitability for drug development, making them ideal choices for our focused properties.

For this task, as introduced in Section 4.5, we combine the property latent embeddings of a reference molecule with another molecule’s context latent embeddings to generate new molecules. To ensure a fair comparison and assess generalization, we use the properties of molecules in the test set as the target when generating new molecules for all baseline methods and our model.

De-novo Molecule Generation. We evaluate the mean squared error (MSE) and mean absolute error (MAE) between the input properties and the real properties of the generated 3D molecules. The results are shown in Table 1, and the detailed standard deviation of experiment results are provided in Appendix 11.3. The baseline results of Asphericity and QED on the GEOM-Drugs dataset are taken from HierDiff (Qiang et al. 2023) while the others are reproduced by ourselves using their official implementation. The results show that our model achieves the best performance on 7 out of 8 properties across two datasets. While we do not achieve the top ranking in the logP task, we attain results that closely approximate the current state-of-the-art method. The results demonstrate the effectiveness of our method in property-targeting generation for 3D drug-like molecules, showing that our model can serve as a useful tool in computational drug design. We provide **results for general generation quality**, i.e., drug-likeness metrics and 3D conformation quality metrics, in Appendix 11.4.

Structure-based Drug Design. In real-world drug design applications, optimizing specific properties of binding molecules is a pivotal scenario. For instance, enhancing the synthetic accessibility score of a binding ligand can preserve its binding capability while simplifying the manufacturing process. Thus, we propose a property-targeting generation task for structure-based drug design. For this task, we train the model by taking pocket-ligand complex structures as input and performing ligand generation conditioned on a given property latent vector. Through this method, we generate 3D molecules while considering specific target binding sites and desired property values. We evaluate the generated molecules from targeted binding affinity and molecular properties. Following previous works (Guan et al. 2023; Luo et al. 2021), we employ AutoDock Vina (Eberhardt et al. 2021) to estimate the target binding affinity. We use the MSE and MAE between generated and reference properties as property-related metrics. The results of are shown in Table 2. The results demonstrate that our model excels in generating binding ligands for target proteins with enhanced property values, surpassing the performance of the baseline model. This underscores the practical utility of our model in real-world structure-based drug design.

Multi-property Targeting Generation. Additionally, we apply our model to disentangle *multiple* target properties and the structure context. Specifically, we conduct experiments with asphericity and QED properties using the GEOM-drug dataset. The results are shown in the following table. We can observe that the performance of multi-property targeting generation slightly dropped. Theoretically, asphericity and QED cannot be completely disentangled, but our model can still achieve a certain degree of explicit control on the multi-property-targeting generation.

5.2 Context-Preserving 3D Generation

Our model enables a novel *structural context-preserving 3D generation* task, which optimizes targeted molecular properties while preserving the molecule’s fundamental architecture. This capability plays a crucial role in enhancing the efficacy,

Table 2: Results on property-guided ligand generation for target protein on CrossDocked2020 dataset. We reported the Vina Dock score for binding affinity and MSE and MAE between reference and actual properties of generated 3D molecules.

	Asphericity			QED			SAS			logP		
	Vina	MSE	MAE	Vina	MSE	MAE	Vina	MSE	MAE	Vina	MSE	MAE
HierDiff	-4.253	0.125	0.296	-4.429	0.113	0.275	-5.053	2.364	1.620	-4.293	5.938	1.855
TargetDiff	-5.742	0.117	0.288	-5.706	0.112	0.307	-5.479	3.369	1.604	-5.501	4.509	1.946
Ours	-5.891	0.102	0.271	-5.866	0.086	0.247	-5.940	2.358	1.529	-5.827	4.376	1.794

Table 3: Results on multi-property targeting generation with asphericity and QED properties on GEOM-drug.

	Asphericity		QED	
	MSE	MAE	MSE	MAE
EDM	0.594	0.432	0.147	0.319
HierDiff	0.193	0.412	0.136	0.301
E3WAE	0.106	0.251	0.099	0.269

safety, and synthesizability of molecules in drug development. For this task, as introduced in Section 4.5, we fuse the context latent embeddings of a template molecule with another molecule’s property latent embeddings to generate new molecules. The structural similarity between the template and the newly generated molecules is quantified by assessing the cosine similarity of their structural latent embeddings.

Retrieval-based Baselines. As no previous method can be adapted to introduce a context condition and perform this context-preserving generation task, we design our baselines via retrieval and evaluate the metrics in the training set. Implementation details are in Appendix 11.2.

The embedding and fingerprint similarity scores between generated molecules and those obtained from retrieval-based baselines are presented in Table 4. The scores for molecules generated by our model surpass nearly all of the mean retrieval similarity scores and, in some cases, even exceed the maximum retrieval values. This performance indicates that our model can generate new molecules that not only possess optimized properties but also closely resemble the reference template structure. Overall, this indicates the effectiveness of our model in context-preserving generation.

5.3 Ablation Study

We perform ablation studies to justify our model design on two key components of E3WAE: (1) model architecture and disentangled representation learning objectives, comparing

Table 4: Results on context-preserving generation. The scores indicate the similarity of context latent variables (Emb.) and fingerprints (Fp.) between reference molecules and molecules generated with our model or retrieved from the training set. The top two results are highlighted as **1st** and **2nd**.

		Asphericity		QED		SAS		logP	
		Emb.	Fp.	Emb.	Fp.	Emb.	Fp.	Emb.	Fp.
GEOM	<i>Ret. Mean</i>	0.33	0.241	0.403	0.243	0.472	0.240	0.437	<u>0.212</u>
	<i>Ret. Max</i>	0.584	0.317	0.821	0.320	0.958	0.318	0.872	0.311
	EDM	0.473	0.130	0.357	0.135	0.396	0.126	0.499	0.131
	HierDiff	0.419	0.148	0.391	0.133	0.514	0.143	0.463	0.161
	Ours	0.593	<u>0.255</u>	<u>0.618</u>	0.218	<u>0.532</u>	<u>0.238</u>	<u>0.624</u>	0.194
Cross-Docked 2020	<i>Ret. Mean</i>	0.260	0.206	0.331	0.204	<u>0.496</u>	0.202	0.503	0.202
	<i>Ret. Max</i>	<u>0.525</u>	<u>0.280</u>	0.566	<u>0.271</u>	0.974	0.273	<u>0.726</u>	<u>0.276</u>
	EDM	0.468	0.152	0.507	0.140	0.378	0.131	0.495	0.147
	HierDiff	0.461	0.185	<u>0.587</u>	0.197	0.412	<u>0.305</u>	0.484	0.200
	Ours	0.570	0.344	0.665	0.307	0.488	0.416	0.763	0.408

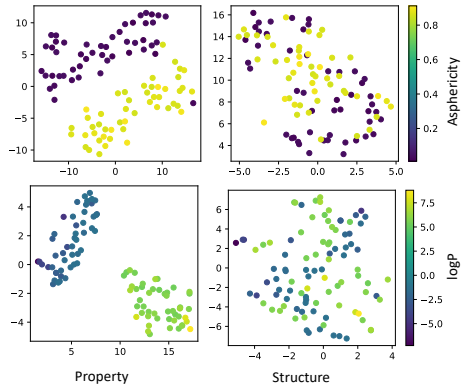


Figure 2: t-SNE visualization of the model’s disentangled latent spaces, colored by ground-truth property values.

WAE and VAE-based frameworks, and (2) the impact of using or omitting our proposed coordinate loss. The results demonstrate diminished performance or delayed convergence with alternatives to our model’s design. Detailed results and analysis are in Appendix 12.

5.4 Visualization of Disentangled Latent Space

To more explicitly show the disentangling ability of our model, we provide a t-SNE visualization (Van der Maaten and Hinton 2008) of the property and structure embeddings in Figure 2, focusing on asphericity and logP as examples. The color coding corresponds to various ground truth property values. We observe that for each molecule, property labels can be well-represented by the property latent variables but not the context latent embeddings. The observed entanglement in the context latent space reflects the inherent independence of structural aspects from selected property labels, in line with the principles of disentangled representation learning. This distinct separation in the property space demonstrates our model’s ability to differentiate molecular properties and align them systematically, thereby enabling explicit control over generating 3D molecules with desired characteristics.

6 Conclusion

We present E3WAE, an E(3)-equivariant Wasserstein autoencoder for conditional generation of 3D molecules with explicit control. Extensive experiments on both over *de-novo* 3D molecule generation and structure-based drug design validate our model’s effectiveness in both property-guided and structure-guided molecule generation. Additionally, our model can be adapted for explicit control over other drug design cases, such as and protein sequence and structure co-design. We leave these directions as future exploration.

Acknowledgement

We would like to thank Limei Wang for insightful discussions and feedback regarding the datasets used in this work.

References

- Adams, K.; and Coley, C. W. 2023. Equivariant Shape-Conditioned Generation of 3D Molecules for Ligand-Based Drug Design. In *The Eleventh ICLR*.
- Axelrod, S.; and Gomez-Bombarelli, R. 2022. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1): 185.
- Barber, D.; and Agakov, F. 2004. The im algorithm: a variational approach to information maximization. *NeurIPS*, 16(320): 201.
- Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; and Csányi, G. 2022. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *NeurIPS*, 35: 11423–11436.
- Bengio, E.; Jain, M.; Korablyov, M.; Precup, D.; and Bengio, Y. 2021. Flow network based generative models for non-iterative diverse candidate generation. *NeurIPS*, 34: 27381–27394.
- Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.
- Brown, N.; Fiscato, M.; Segler, M. H.; and Vaucher, A. C. 2019. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3): 1096–1108.
- Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β -VAE. *arXiv preprint arXiv:1804.03599*.
- Ceylan, C.; and Gutmann, M. U. 2018. Conditional noise-contrastive estimation of unnormalised models. In *ICML*, 726–734. PMLR.
- Chen, X.; Duan, Y.; Houthoofd, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NeurIPS*, 29.
- Cheng, P.; Min, M. R.; Shen, D.; Malon, C.; Zhang, Y.; Li, Y.; and Carin, L. 2020. Improving Disentangled Text Representation Learning with Information-Theoretic Guidance. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7530–7541. Online: Association for Computational Linguistics.
- Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.
- Deng, C.; Litany, O.; Duan, Y.; Poulenard, A.; Tagliasacchi, A.; and Guibas, L. J. 2021. Vector neurons: A general framework for so (3)-equivariant networks. In *ICCV*, 12200–12209.
- Deng, Y.; Bakhtin, A.; Ott, M.; Szlam, A.; and Ranzato, M. 2020. Residual Energy-Based Models for Text Generation. In *ICLR*.
- Denton, E. L.; et al. 2017. Unsupervised learning of disentangled representations from video. *NeurIPS*, 30.
- Du, Y.; Guo, X.; Wang, Y.; Shehu, A.; and Zhao, L. 2022. Small molecule generation via disentangled representation learning. *Bioinformatics*, 38(12): 3200–3208.
- Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; and Forli, S. 2021. AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8): 3891–3898.
- Fey, M.; and Lenssen, J. E. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
- Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; and Koes, D. R. 2020. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9): 4200–4215.
- Fu, T.; Xiao, C.; Li, X.; Glass, L. M.; and Sun, J. 2021. Mimosa: Multi-constraint molecule sampling for molecule optimization. In *AAAI*, volume 35, 125–133.
- Garcia Satorras, V.; Hoogeboom, E.; Fuchs, F.; Posner, I.; and Welling, M. 2021. E (n) equivariant normalizing flows. *NeurIPS*, 34: 4181–4192.
- Gebauer, N.; Gastegger, M.; and Schütt, K. 2019. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *NeurIPS*, 32.
- Geiger, M.; and Smidt, T. 2022. e3nn: Euclidean Neural Networks.
- Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773.
- Guan, J.; Qian, W. W.; Peng, X.; Su, Y.; Peng, J.; and Ma, J. 2023. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. In *The Eleventh ICLR*.
- Han, J.; Min, M. R.; Han, L.; Li, L. E.; and Zhang, X. 2021. Disentangled Recurrent Wasserstein Autoencoder. In *ICLR*.
- Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2016. beta-vaes: Learning basic visual concepts with a constrained variational framework. In *ICLR*.
- Hoogeboom, E.; Satorras, V. G.; Vignac, C.; and Welling, M. 2022. Equivariant diffusion for molecule generation in 3d. In *ICML*. PMLR.
- Huang, H.; Sun, L.; Du, B.; and Lv, W. 2023. Learning Joint 2D & 3D Diffusion Models for Complete Molecule Generation. *arXiv preprint arXiv:2305.12347*.
- Huang, Y.; Peng, X.; Ma, J.; and Zhang, M. 2022. 3DLinker: an E (3) equivariant variational autoencoder for molecular linker design. *arXiv preprint arXiv:2205.07309*.
- Igashov, I.; Stärk, H.; Vignac, C.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; and Correia, B. 2022. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274*.

- Imrie, F.; Bradley, A. R.; van der Schaar, M.; and Deane, C. M. 2020. Deep generative models for 3D linker design. *Journal of chemical information and modeling*, 60(4): 1983–1995.
- Jin, W.; Barzilay, R.; and Jaakkola, T. 2018. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *ICML*, 2323–2332.
- Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5): 922–923.
- Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *ICML*, 2649–2658. PMLR.
- Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd ICLR, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Kotovenko, D.; Sanakoyeu, A.; Lang, S.; and Ommer, B. 2019. Content and style disentanglement for artistic style transfer. In *ICCV*, 4422–4431.
- Landrum, G. 2016. RDKit: open-source cheminformatics <http://www.rdkit.org>. 3(8).
- Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, 35–51.
- Li, T.; Guo, H.; Grazioli, F.; Gerstein, M.; and Min, M. R. 2023. Disentangled Wasserstein Autoencoder for T-Cell Receptor Engineering. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Liu, M.; Luo, Y.; Uchino, K.; Maruhashi, K.; and Ji, S. 2022. Generating 3D Molecules for Target Protein Binding. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th ICML*, volume 162 of *Proceedings of Machine Learning Research*, 13912–13924. PMLR.
- Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 4114–4124. PMLR.
- Luo, S.; Guan, J.; Ma, J.; and Peng, J. 2021. A 3D generative model for structure-based drug design. *NeurIPS*.
- Luo, Y.; and Ji, S. 2022. An Autoregressive Flow Model for 3D Molecular Geometry Generation from Scratch. In *ICLR*.
- Mollaysa, A.; Paige, B.; and Kalousis, A. 2020. Conditional generation of molecules from disentangled representations.
- Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.
- Polishchuk, P.; G, T. I.; Madzhidov; and Varnek, A. 2013. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design*, 27: 675–679.
- Powers, A. S.; Yu, H. H.; Suriana, P. A.; and Dror, R. O. 2022. Fragment-based ligand generation guided by geometric deep learning on protein-ligand structures. In *ICLR2022 Machine Learning for Drug Discovery*.
- Qiang, B.; Song, Y.; Xu, M.; Gong, J.; Gao, B.; Zhou, H.; Ma, W.-Y.; and Lan, Y. 2023. Coarse-to-fine: a hierarchical diffusion model for molecule generation in 3d. In *ICML*, 28277–28299. PMLR.
- Satorras, V. G.; Hoogeboom, E.; and Welling, M. 2021. E(n) equivariant graph neural networks. In *ICML*, 9323–9332. PMLR.
- Shi, C.; Xu, M.; Zhu, Z.; Zhang, W.; Zhang, M.; and Tang, J. 2019. GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation. In *ICLR*.
- Sutherland, D. J.; Tung, H.-Y.; Strathmann, H.; De, S.; Ramdas, A.; Smola, A.; and Gretton, A. 2017. Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy. In *ICLR*.
- Tolstikhin, I.; Bousquet, O.; Gelly, S.; and Schoelkopf, B. 2018. Wasserstein Auto-Encoders. In *ICLR*.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).
- Xie, Y.; Shi, C.; Zhou, H.; Yang, Y.; Zhang, W.; Yu, Y.; and Li, L. 2021. {MARS}: Markov Molecular Sampling for Multi-objective Drug Discovery. In *ICLR*.
- Xu, M.; Powers, A. S.; Dror, R. O.; Ermon, S.; and Leskovec, J. 2023. Geometric latent diffusion models for 3d molecule generation. In *ICML*, 38592–38610. PMLR.
- Yang, S.; Hwang, D.; Lee, S.; Ryu, S.; and Hwang, S. J. 2021. Hit and lead discovery with explorative rl and fragment-based molecule generation. *NeurIPS*, 34: 7924–7936.
- Yu, R. 2020. A Tutorial on VAEs: From Bayes’ Rule to Lossless Compression. *arXiv preprint arXiv:2006.10273*.
- Zhu, J.-Y.; Zhang, Z.; Zhang, C.; Wu, J.; Torralba, A.; Tenenbaum, J.; and Freeman, B. 2018. Visual object networks: Image generation with disentangled 3D representations. *NeurIPS*, 31.
- Zhu, Y.; Min, M. R.; Kadav, A.; and Graf, H. P. 2020. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *CVPR*, 6538–6547.

7 Fragmentizing the Molecules

To streamline the modeling of large molecules, fragment-based generation has been favored over atomic models, prompting the development of methods to segment molecules into fragments. An optimal decomposition algorithm should yield a comprehensive yet concise fragment vocabulary.

JT-VAE (Jin, Barzilay, and Jaakkola 2018) is the pioneer work for fragment-based molecule generation, using a minimum spanning tree algorithm to decompose molecules while maintain chemical bonds and avoid cycles. It achieved coverage of all synthetically accessible structures with a vocabulary under 800. Subsequent models like MARS (Xie et al. 2021), FREED (Yang et al. 2021), and MIMOSA (Fu et al. 2021), each develop unique bond-breaking criteria, and further refines the vocabulary while ensuring frequency-based relevance. A recent study, HierDiff (Qiang et al. 2023), expands upon this approach by leveraging the tree decomposition algorithm from JT-VAE for fragment-based 3D molecule generation, introducing center coordinates as the three-dimensional representation for fragments.

Given the vast chemical space, it is practical to exclude less common fragments. Following HierDiff (Qiang et al. 2023), we employ Jin, Barzilay, and Jaakkola (2018) tree decomposition in 3D space, which includes the following steps: Firstly, we extract chemical bonds not forming part of any rings and identify isolated rings representing single topological cycles. Secondly, we merge bridged rings that share over two atoms, acknowledging their chemical importance and distinct 3D conformations. Thirdly, to avoid cyclic structures, any intersecting atom bonded to more than three others is incorporated into the graph as an individual fragment. Finally, we provide each fragment with a 3D geometric representation based on the centroid of its atoms. By this fragmentizing method, we balance structural complexity with the need for a reasonable-sized fragment vocabulary. For the encoder, we use the high-level features based on property values as outlined in HierDiff, along with the embeddings of fragment types, as the input features for each fragment.

8 Disentanglement Guarantee

In this section, we provide a theoretical analysis to demonstrate our model’s capacity for guaranteeing the disentanglement of property and context latent variables. Following previous work on disentangled representation learning (Cheng et al. 2020; Han et al. 2021; Li et al. 2023), we define a measurement of embedding disentanglement as follows,

$$D(\mathbf{z}_p, \mathbf{z}_s; \mathbf{x}) = \text{VI}(\mathbf{z}_p; \mathbf{x}) + \text{VI}(\mathbf{z}_s; \mathbf{x}) - \text{VI}(\mathbf{z}_p; \mathbf{z}_s), \quad (15)$$

where \mathbf{z}_p and \mathbf{z}_s are the property and context latent variables, respectively, and \mathbf{x} is the input molecule. Variation of Information, VI, is a measure of independence between two random variables. The mathematical definition of VI between random variables \mathbf{x} and \mathbf{y} is

$$\text{VI}(\mathbf{x}; \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - 2I(\mathbf{x}; \mathbf{y}), \quad (16)$$

where $H(\mathbf{x})$ and $H(\mathbf{y})$ are entropies of \mathbf{x} and \mathbf{y} , respectively, and $I(\mathbf{x}; \mathbf{y})$ is the mutual information (MI) between variables \mathbf{x} and \mathbf{y} . Additionally, $\text{VI}(\mathbf{x}; \mathbf{y}) = 0$ indicates \mathbf{x} and \mathbf{y} are the sample variable. Intuitively, the mutual information is a measure of “dependence” between two variables, and the VI distance is a measure of “independence” between them.

This measurement reaches 0 when \mathbf{z}_p and \mathbf{z}_s are totally independent, *i.e.*, fully disentangled. Thus, we treat this measurement as our disentangling objective.

By the definition of VI, the measurement D can be simplified as:

$$\text{VI}(\mathbf{z}_p; \mathbf{x}) + \text{VI}(\mathbf{z}_s; \mathbf{x}) - \text{VI}(\mathbf{z}_p; \mathbf{z}_s) = 2H(\mathbf{x}) + 2[I(\mathbf{z}_p; \mathbf{z}_s) - I(\mathbf{z}_p; \mathbf{x}) - I(\mathbf{z}_s; \mathbf{x})]. \quad (17)$$

Since $H(\mathbf{x})$ is a constant associated with the data, we only need to focus on $I(\mathbf{z}_p; \mathbf{z}_s) - I(\mathbf{z}_p; \mathbf{x}) - I(\mathbf{z}_s; \mathbf{x})$.

Giving rise to the problem that without any inductive bias in supervision, the disentangled representation could be meaningless (Locatello et al. 2019), we add inductive biases by using the property value y as supervised information for the property latent \mathbf{z}_p . As $\mathbf{z}_p \rightarrow \mathbf{x} \rightarrow y$ forms a Markov chain, we have $I(\mathbf{z}_p; \mathbf{x}) \geq I(\mathbf{z}_p; y)$ based on the MI data-processing inequality (Cover 1999). Therefore, we derive an upper bound for the measurement as:

$$I(\mathbf{z}_p; \mathbf{z}_s) - I(\mathbf{z}_p; \mathbf{x}) - I(\mathbf{z}_s; \mathbf{x}) \leq I(\mathbf{z}_p; \mathbf{z}_s) - I(\mathbf{z}_p; y) - I(\mathbf{z}_s; \mathbf{x}). \quad (18)$$

Next, we show how our model design and training objective aligns to the optimization of the terms $I(\mathbf{z}_p; \mathbf{z}_s)$, $I(\mathbf{z}_p; y)$, and $I(\mathbf{z}_s; \mathbf{x})$ in Equation 18.

Minimizing $I(\mathbf{z}_p; \mathbf{z}_s)$. By minimizing the Wasserstein regularization term \mathcal{L}_{Dis} , we forces the distribution of the combined latent variables $\mathbf{z} = \text{concat}(\mathbf{z}_p, \mathbf{z}_s)$ to approach an isotropic Gaussian distribution, where all dimensions are independent. This enforces the independence across all dimensions of \mathbf{Z} . Consequently, it minimizes the mutual information between the latent variables \mathbf{z}_p and \mathbf{z}_s , aiding in their disentanglement.

Maximizing $I(\mathbf{z}_p; y)$. By leveraging an MI variational lower bound (Barber and Agakov 2004), we derive a lower bound for $I(\mathbf{z}_p; y)$ as

$$I(\mathbf{z}_p; y) \geq H(y) + \mathbb{E}_{\pi(y, \mathbf{z}_p)} [\log q_{\mathcal{H}}(y|\mathbf{z}_p)]. \quad (19)$$

Concretely, the latent variable \mathbf{z}_p is encouraged to give a better prediction of property value y by maximizing $\mathbb{E}_{\pi(y, \mathbf{z}_p)} [\log q_{\mathcal{H}}(y|\mathbf{z}_p)]$, where $q_{\mathcal{H}}(y|\mathbf{z}_p)$ is the predicted probability by the auxiliary prediction head \mathcal{H} . Thus, optimizing the performance of \mathcal{H} maximizes $I(\mathbf{z}_p; y)$.

Maximizing $I(\mathbf{z}_s; \mathbf{x})$. According to Theorem 3 in Han et al. (2021) and Theorem 2.1 in Li et al. (2023), given the encoder $Q_{\theta}(\mathbf{z}|\mathbf{x})$, prior $P_{\mathbf{z}}$ and the data distribution P_D , we have

$$\mathbb{D}_{\text{KL}}(Q_{\mathbf{z}}\|P_{\mathbf{z}}) = \mathbb{E}_{P_D} [\mathbb{D}_{\text{KL}}(Q_{\theta}(\mathbf{z}|\mathbf{x})\|P_{\mathbf{z}})] - I(\mathbf{z}; \mathbf{x}), \quad (20)$$

where KL represents the Kullback-Leibler divergence and the $Q_{\mathbf{z}}$ is the marginal distribution of the encoder when $\mathbf{x} \sim P_D$ and $\mathbf{z} \sim Q_{\theta}(\mathbf{z}|\mathbf{x})$. Equation 20 demonstrate that by minimizing the KL divergence between $Q_{\mathbf{z}}$ and $P_{\mathbf{z}}$, we jointly maximize the mutual information between the data \mathbf{x} and the latent variable \mathbf{z} . This also applies to the two separate parts of \mathbf{z} , \mathbf{z}_p and \mathbf{z}_s . In practice, measuring the marginal directly is often infeasible. Consequently, we resort to minimizing the kernel MMD outlined in Equation 5. This approach obviates the necessity for additional constraints on the information content of \mathbf{z}_s , as $I(\mathbf{z}_s; \mathbf{x})$ is inherently maximized by the objective.

9 3D Molecule Graph Reconstruction Algorithm

The 3D molecule graph reconstruction process, i.e., the decoding process, is illustrated in Figure 3 Algorithm 1.

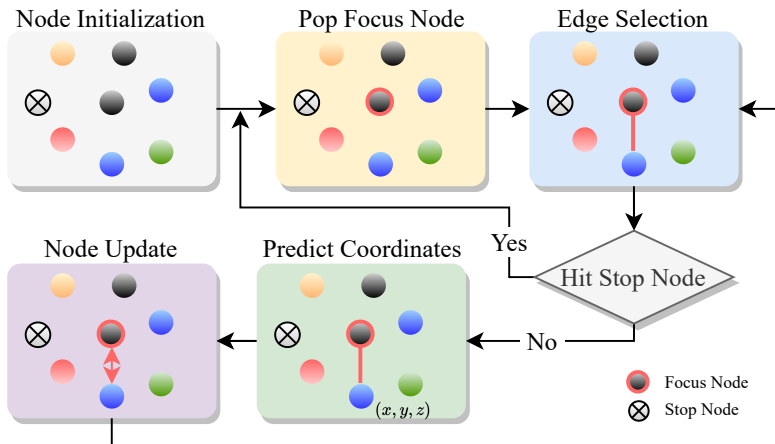


Figure 3: An illustration of the reconstruction process.

10 Dataset Description

GEOM-Drugs GEOM-Drugs (Axelrod and Gomez-Bombarelli 2022) is a large-scale dataset that contains molecules with up to 181 atoms and an average of 44.4 atoms. GEOM-Drugs provides multiple conformations for each molecule with corresponding energies, and we retain one stable conformation with the lowest energy to construct the dataset. We take the first 50k molecules to train our model. The training/validation/test split ratio is 70%, 15%, and 15%, respectively.

CrossDocked2020 CrossDocked2020 (Francoeur et al. 2020) is a dataset designed for the development and benchmarking of structure-based drug discovery. It comprises 22.5 million poses of ligands docked into a variety of similar binding pockets sourced from the Protein Data Bank. Our study focus on the ligand structures and corresponding drug-like properties derived from this dataset. For model training, we select a subset of 50k molecules. The training/validation/test split ratio is 70%, 15%, and 15%, respectively.

Algorithm 1: 3D Molecule Graph Reconstruction in E3WAE

```
1: Input: Latent variables  $\mathbf{z}_h, \mathbf{z}_v$ 
2: Initialize:
    $\{x_i\}_{i=1,\dots,n} \leftarrow \text{NodeTypes}(\mathbf{z}_h, \mathbf{z}_v)$ 
   Queue  $Q \leftarrow \emptyset$ 
    $Q.\text{push}(\text{RandomSelect}(\{1, \dots, n\}))$ 
   3D Graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{R}\}$ , where  $\mathcal{V} \leftarrow \emptyset, \mathcal{E} \leftarrow \emptyset, \mathcal{R} \leftarrow \emptyset$ 
3: while  $Q \neq \emptyset$  do
4:    $f \leftarrow Q.\text{pop}()$ 
5:   Add node  $f$  to graph  $\mathcal{G}$ 
6:   isStopNode  $\leftarrow$  False
7:   while not isStopNode do
8:      $(i, \text{isStopNode}) \leftarrow \text{PredictEdge}(f, \mathcal{G})$ 
9:     if not isStopNode and FirstLink( $i$ ) then
10:       $\mathbf{r}_u = (x_u, y_u, z_u) \leftarrow \text{PredictCoords}(i)$ 
11:       $\mathcal{V} \leftarrow \mathcal{V} \cup \{i\}, \mathcal{E} \leftarrow \mathcal{E} \cup \{(f, i)\},$ 
12:       $\mathcal{R} \leftarrow \mathcal{R} \cup \{\mathbf{r}_i\}$ 
13:       $Q.\text{push}(i)$ 
14:     end if
15:     MarkVisited( $f$ )
16:   end while
17: end while
18: Return: Reconstructed  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ 
```

11 Additional Experimental Details

This section describes the full experiment setup in this paper. The implementation of our methods is based on the PyTorch (Paszke et al. 2017) and Pytorch Geometric (Fey and Lenssen 2019), and all models are trained with the Adam optimizer (Kingma and Ba 2015). All experiments are conducted on a single NVIDIA A100 80GB GPU. The search space for model and training hyperparameters are listed in Table 5. Note that we select hyperparameters for experiments on asphericity, QED, SA Score, logP properties by the same search space, and the optimal hyperparameters are chosen by the performance on the validation set.

Table 5: Model and training hyperparameters for our method on different datasets.

Hyperparameter	Values/Search Space	
	GROM-Drugs	CrossDocked2020
Number of layers	4, 5, 6, 7, 8, 9	4, 6, 9
Hidden dim	28, 32	28, 32
Coordinate loss trade-off weight	0.5, 1, 3	0.5, 1, 3
Property prediction loss trade-off weight	3, 5, 10, 15	5, 10, 15
Wasserstein loss trade-off weight	3, 5, 10	5, 10, 15
Weight Decay	1e-4	1e-4
Epochs	50	50
Batch size	16, 32	16, 32
Learning rate	1e-4, 2e-4, 5e-4, 1e-3	1e-4, 2e-4, 5e-4
Learning rate decay factor	0.1, 0.5	0.1, 0.5
Learning rate decay epochs	3, 5, 10	3, 5, 10, 20

11.1 Implementation of Baselines

EDM. For EDM (Hoogeboom et al. 2022), we use the official implementation available on GitHub¹. We adhered to the default hyperparameters specified in the repository. To facilitate property-targeting generation, we train EDM models to incorporate drug-like property values as an additional input, concatenated to the node features. During the generation phase, we sample the number of nodes from the node distribution of the training set and employ property values from the test set as reference for generating new 3D molecules.

¹https://github.com/ehoogeboom/e3_diffusion_for_molecules, The MIT License.

HierDiff. We implement HierDiff (Qiang et al. 2023) using its official GitHub repository², following the prescribed default hyperparameters. For property-targeting generation, we trained both the coarse-grained diffusion and edge denoising generation models, integrating drug-like property values into the node features as context. During generation, the fragment numbers was determined based on the distribution observed in the training set, with property values from the test set guiding the generation of 3D molecules.

TargetDiff. We implement TargetDiff (Guan et al. 2023) with its official GitHub repository³, following the prescribed default hyperparameters. For adapting the original work to property-targeting generation, we use a straightforward extension similar to conditional EDM (Hooeboom et al. 2022) Section 3.4 and Appendix E. Specifically, a conditioning on a property value c is added where relevant. The diffusion process that adds noise is not altered. The generative denoising process is conditioned on c by adding it as input to the neural network.

Note that, for context-preserving generation, the results of EDM and HierDiff in Section 5.2 are calculated by taking a molecule with same property values as the reference, since they cannot take the template structure as guidance for conditional generation.

11.2 Implementation of Retrieval-based Baselines

Specifically, we first encode and save the context latent variables in the training set using our E3WAE structure encoder Θ_s . Subsequently, we retrieve molecules that exhibit property values within a specified threshold. For comparison, we calculate and report both the mean (*Ret. Mean*) and the maximum (*Ret. Max*) similarity of structural embeddings between the template molecule and the molecules retrieved from the training set. This similarity metric serves as an indicator of structural resemblance, demonstrating the capability of our model in this task.

11.3 Standard deviation on property-targeting generation

To quantify the variability and reliability of our results, we present the standard deviation of results for property-targeting generation in Table 6.

Table 6: Standard deviation on property-targeting generation.

		Asphericity		QED		SAS		logP	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GEOM-Drugs	EDM	0.088	0.153	0.087	0.153	6.317	0.965	7.427	1.442
	HierDiff	0.096	0.166	0.124	0.186	1.908	0.759	2.531	0.855
	Ours	0.075	0.130	0.071	0.134	2.215	0.747	4.900	1.241
CrossDocked2020	EDM	0.083	0.150	0.131	0.179	6.368	0.983	7.199	1.372
	HierDiff	0.082	0.147	0.069	0.132	1.567	0.645	8.348	1.468
	Ours	0.095	0.123	0.073	0.142	1.775	0.701	6.008	1.241

11.4 General Generation Quality

In this section, we evaluate and report the general generation quality of our model, focusing on two key aspects: drug-likeness and 3D conformation quality.

Drug-Likeness Evaluation. The drug-likeness of generated molecules is a critical measure of their potential as drug candidates. Following the experiment settings in Qiang et al. (2023), we evaluate our model using additional metrics: Retrosynthetic Accessibility (RA), Medicinal Chemistry Filter (MCF) (Brown et al. 2019), $\Delta\log P$, and ΔMW . RA assesses the synthetic feasibility of molecules, while MCF quantifies the percentage of molecules devoid of undruggable substructures. $\Delta\log P$ measures the deviation of logP values from the ground truth of the training set, and ΔMW evaluates the variance in Molecular Weight from the training set’s standard. Our model shows competitive performance in these metrics, particularly excelling in MCF scores and showing close proximity to SOTA methods in others. Moreover, it’s important to mention that these baseline methods typically require the input of a node number derived from the node distribution, a condition that is often impractical in real-world scenarios. These results demonstrate the model’s capability to effectively generate drug-like molecules.

Conformation Quality Evaluation. The evaluation of 3D generation quality is conducted using stability metrics and substructure geometry alignment metrics. Stability metrics, as introduced by Hooeboom et al. (2022), includes atomic stability (the proportion of atoms with correct valency) and molecular stability (the proportion of molecules where all atoms are stable).

²<https://github.com/qiangbo1222/HierDiff>, The MIT License.

³<https://github.com/guanjq/targetdiff>, The MIT License.

Table 7: Results on drug-likeness evaluation, where the last column stands for the results of training data. * denotes methods that contain a diffusion model and need to take a node number from the node distribution as input. GSNet is short for G-SphereNet. Experiments that report errors due to low validity are marked as ‘-’.

	GSNet	EDM*	HierDiff*	Ours	GEOM-Drugs
QED \uparrow	0.382	0.505	0.509	0.527	0.658
RA \uparrow	--	0.368	0.611	0.559	0.915
MCF \uparrow	0.489	0.552	0.654	0.763	0.774
SAS \downarrow	--	5.346	4.500	4.077	4.018
$\Delta\log P \downarrow$	2.306	0.577	0.279	0.194	0.000
$\Delta MW \downarrow$	170.700	24.578	19.781	69.085	0.000

	GSNet	EDM*	HierDiff*	Ours	CrossDocked
QED \uparrow	0.442	0.402	0.486	0.435	0.619
RA \uparrow	--	0.358	0.585	0.716	0.912
MCF \uparrow	0.449	0.416	0.726	0.645	0.746
SAS \downarrow	--	7.325	4.536	4.328	2.564
$\Delta\log P \downarrow$	3.359	1.576	1.132	0.851	0.000
$\Delta MW \downarrow$	200.950	28.856	4.763	76.801	0.000

Table 8: Results on 3D conformation generation quality evaluation, with metrics for molecular and atomic stability, as well as MMD for bond, angle, and dihedral angles.

	Geom-Drugs			CrossDocked2020		
	EDM	HierDiff	Ours	EDM	HierDiff	Ours
Mol. Sta. (%)	0.1	31.7	25.6	0.2	35.7	50.9
Atom Sta. (%)	81.2	88.4	79.9	73.1	83.1	87.3
Bond MMD	1.034	0.989	0.712	0.487	0.580	0.341
Angle MMD	1.200	0.871	0.916	0.562	0.151	0.386
Dihedral MMD	0.028	0.082	0.070	0.040	0.051	0.094

For the substructure geometry alignment metrics, following Huang et al. (2023), we select the 8 most frequent types of bonds, bond pairs, and bond triples, and compute the MMD distances on bond lengths, bond angles, and dihedral angles distributions between generated molecules and test set. As shown in Table 8, the performance of our model either surpass or are on par with those of diffusion-based methods, known for their high computational complexity. This indicates the model’s effective generation of molecules with not only realistic but also structurally stable 3D conformations.

12 Ablation Studies

12.1 Ablation Study on Disentangle Representation Learning Framework

To elucidate the rationale behind our model’s design, we incorporate two VAE baselines E3fVAE and E3dVAE, inspired by FactorVAE (Kim and Mnih 2018) and IDEL’s objectives from Cheng et al. (2020).

The first baseline, E3fVAE, adds a total correlation term for enhanced disentanglement without compromising reconstruction quality. We conduct experiments adapting the disentangle objective in FactorVAE (Kim and Mnih 2018) for 3D molecule generation referencing the official implementation on GitHub⁴. Specifically, we use a 3-layer EGNN as the discriminator and the training objectives are:

- The reconstruction loss for 3D molecule graphs: \mathcal{L}_{Recon}
- The KL divergence term for VAE $\mathcal{L}_{KL} = \mathbb{D}_{KL}(Q_{\theta}(\mathbf{z}|\mathbf{x})||P_{\mathbf{z}})$, where $\mathbf{z} = \text{concat}(\mathbf{z}_p, \mathbf{z}_s)$
- The total correlation term
- The discriminator loss to approximate the density ratio in KL term

The second baseline, E3dVAE, focuses on minimizing a mutual information (MI) upper bound, a strategy aimed at achieving disentanglement within the embedding space. For training, the loss comprises of the following components:

- The reconstruction loss for 3D molecule graphs: \mathcal{L}_{Recon}

⁴<https://github.com/1Konny/FactorVAE>. The MIT license.

- The KL divergence term for VAE: $\mathcal{L}_{\text{KL}} = \mathbb{D}_{\text{KL}}(Q_{\theta}(\mathbf{z}|\mathbf{x})||P_{\mathbf{z}})$, where $\mathbf{z} = \text{concat}(\mathbf{z}_p, \mathbf{z}_s)$
- The reconstruction loss given \mathbf{z}_s : $\mathcal{L}_{\text{Recon},s}(\mathcal{G}, \mathcal{D}_s(\hat{\mathcal{G}}|\mathbf{z}_s))$, where \mathcal{D}_s is an auxiliary decoder to reconstruct the 3D molecule graph \mathcal{G} by the context latent variable \mathbf{z}_s
- The property prediction loss given $\mathbf{z}_{h,p}$: $\mathcal{L}_{\text{prop}}$
- The sample-based MI upper bound between the latent variables: $\mathcal{L}_{\text{MI}}(\mathbf{z}_p, \mathbf{z}_s)$. This requires an approximation of the conditional distribution $p(\mathbf{z}_{h,p}|\mathbf{z}_{h,s})$, which is approximated by a separate variational network.

E3WAE, differing from the above VAE-based models, is deterministic and has simpler objectives, which helps avoid typical VAE training challenges such as hyperparameter sensitivity. We perform an ablation study with GEOM-Drug dataset on property-targeting generation with Asphericity and QED. The results in Table 9 reveal a performance drop in E3dVAE compared to E3WAE. Considering this and the architectural simplicity of E3WAE, we opt for a WAE-based framework, E3WAE, as our preferred model.

Table 9: Ablation study on the architecture choice.

	Asphericity		QED		# Para.
	MSE	MAE	MSE	MAE	
E3fVAE	0.143	0.381	0.112	0.306	3.2M
E3dVAE	0.101	0.260	0.098	0.257	2.4M
E3WAE	0.095	0.246	0.072	0.221	2.1M

12.2 Ablation Study on the Proposed Coordinate Reconstruction Loss

In this ablation study, we assess the impact of the proposed coordinate reconstruction loss on model training. Figure 4 shows the loss curves on training and validation set: with and without the proposed alignment strategy. This indicate that incorporating the alignment into the coordinate loss results in expedited convergence and a lower convergence plateau. This suggests that the alignment component in the loss function plays a critical role in guiding the model to a more accurate and stable solution. Based on these observations, we opt for the proposed coordinate loss in our model to leverage its benefits for more efficient training and improved model performance on validation data.

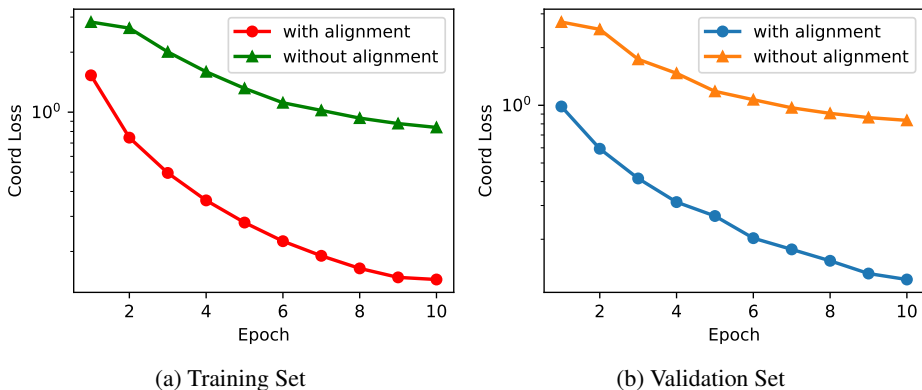


Figure 4: Comparison of training coordinate loss curves: proposed coordinate loss vs. original log-MSE loss without structural alignment.

13 Visualization of Generated Molecules

The visualizations of generated molecules for asphericity-guided generation tasks are shown in Figure 5 for EDM, Figure 6 for HierDiff, and Figure 7 for our proposed E3WAE, respectively. It’s worth noting that while EDM may exhibit unstable or broken substructures, such as unrealistic rings, and HierDiff could produce unusual connections, our model consistently generates high-quality molecules with more stable substructures.

14 Limitations

While our model provides explicit control over molecular properties and structural contexts, the interpretability of the context latent vectors remains a challenge. Supervision through reliable labeled data and human domain prior are required for disentangled representation learning model design. On the other hand, the generalization ability of our model lacks exploration. More consistent and comprehensive evaluations are needed to confirm the robustness and reliability of our approach in specific drug design scenarios and domains.

15 Broader Impacts

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

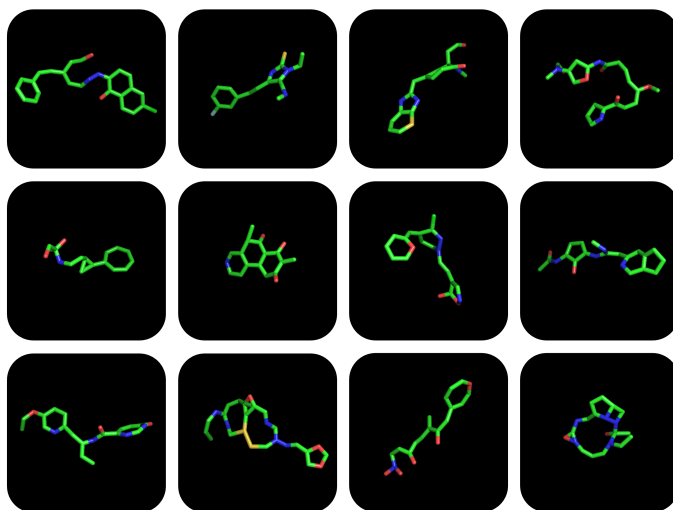


Figure 5: Visualized 3D conformations generated by EDM.

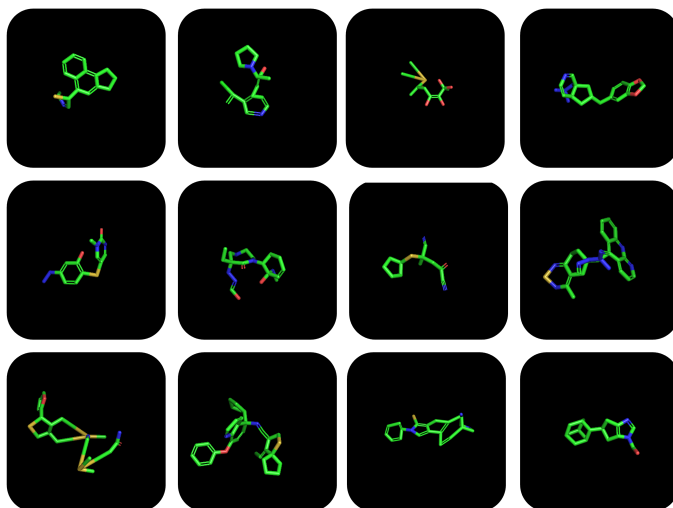


Figure 6: Visualized 3D conformations generated by HierDiff.

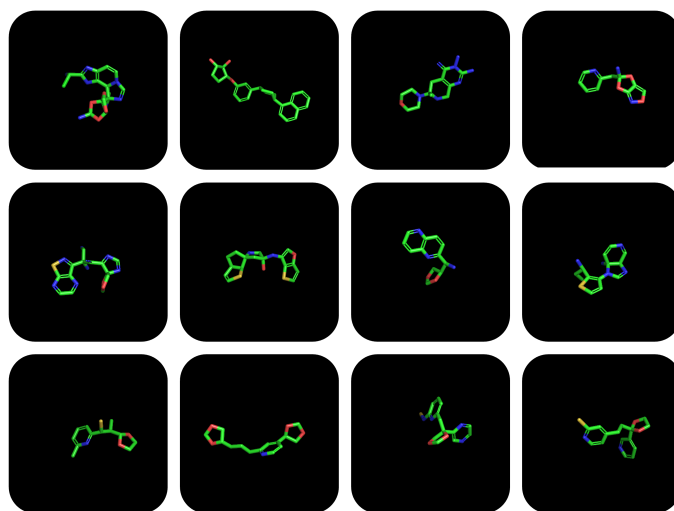


Figure 7: Visualized 3D conformations generated by E3WAE.