Equivariant Blurring Diffusion for Multiscale Generation of Molecular Conformer

Jiwoong Park and Yang Shen
Department of Electrical and Computer Engineering
Texas A&M University
ptywoong@gmail.com, yshen@tamu.edu

Abstract

In this paper, we focus on a fundamental biochemical problem of generating 3D molecular conformers conditioned on molecular graphs in a multiscale manner. Our approach consists of two hierarchical stages: i) generation of coarse-grained 3D structure from the molecular graph, and ii) generation of fine atomic details from the coarse-grained approximated structure. For the challenging second stage, we introduce a novel generative model termed *Equivariant Blurring Diffusion* (EBD), which defines a forward process that moves towards the coarse-grained structure by blurring the fine atomic details of conformers, and a reverse process that performs the opposite operation using equivariant networks. We demonstrate the effectiveness of EBD on a benchmark of drug-like molecules. Codes are released at https://github.com/Shen-Lab/EBD.

1 Introduction

The advancement of generative models to understand the multiscale properties of objects facilitates their application across a range of granularity levels. In the field of biochemistry and drug discovery, however, denoising diffusion models for 3D conformers of stable molecular structures have not yet taken advantage of coarse-to-fine multiscale frameworks. Current methods either disregard the scale hierarchy [44, 55, 18, 22, 56] or consider that in very limited ways [36, 39].

The primary bottleneck in extending denoising diffusion models [46, 47, 49, 14, 25] for molecular conformers to multiscale designs is that random noise corrupts not only fine atomic details but also structural information of coarse-grained structures indiscriminately. To tackle this challenging problem, we exploit *fragments* that are frequently occurring substructures or functional groups in 2D molecular graphs. Introducing fragments divides the generation process into two stages: i) generating coarse-grained structures represented by fragments, and ii) restoring fine atomic details from fragment structures. In the first stage of generating fragment coordinates from molecular graphs, we efficiently generate approximations of fragment structures comprising the center of mass and attributes of each fragment from a cheminformatics tool.

For the challenging second step of coarse-to-fine generation, we propose a novel diffusion model, *Equivariant Blurring Diffusion* (EBD). Motivated from the blurring corruption of the heat equation (IHDM) [40], we design EBD to generate 3D molecular conformers from coarse-grained fragment approximated structures, rather than from random noise. The forward process moves atom positions of conformers towards the center of mass of their respective fragments, while the reverse process restores full-atom details from the prior distribution of the 3D fragment structure. The designed blurring schedule allows the diffusion model to focus on restoring fine atomic details while retaining coarse-grained information throughout the entire generative process. We validated EBD model using a benchmark of drug-like molecules. We obtained superior results in conformer generation compared to the denoising diffusion model, even with 100 times fewer diffusion time steps.

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

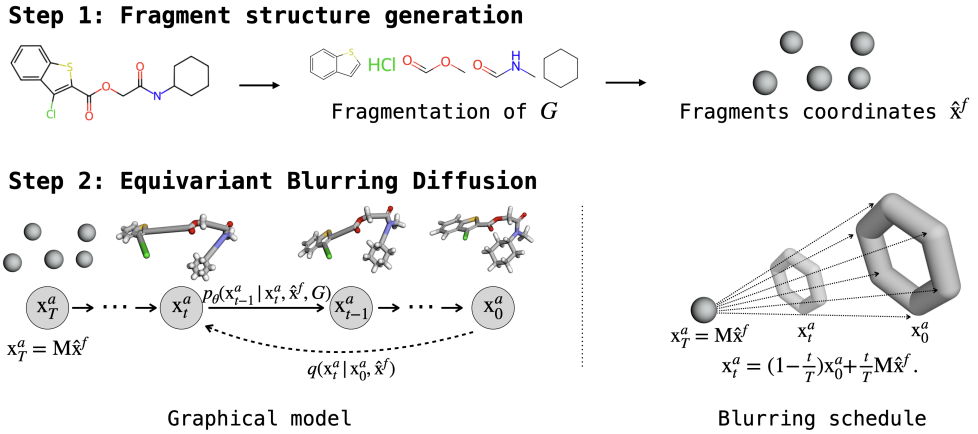


Figure 1: Our hierarchical molecular conformer generation framework. We first decompose a molecular graph G into fragments and generate fragment coordinates $\hat{\mathbf{x}}^f$. Then, conditioned on $\hat{\mathbf{x}}^f$ and G, Equivariant Blurring Diffusion generates atom-level fine details using the blurring schedule.

2 Methods

Our objective is to generate an ensemble of 3D molecular conformers $\mathbf{x}^{\mathrm{a}} \in \mathbb{R}^{n\times 3}$ given a molecular graph G of n atoms. Our hierarchical approach is in two stages. i) $p(\mathbf{x}^{\mathrm{f}}|G)$: generating a coarse-grained 3D structure of fragment coordinates $\mathbf{x}^{\mathrm{f}} \in \mathbb{R}^{m\times 3}$ from G which was decomposed into m fragments, and ii) $p(\mathbf{x}^{\mathrm{a}}|\mathbf{x}^{\mathrm{f}}, G)$: the diffusion model generating fine atomic details $\mathbf{x}^{\mathrm{a}} \in \mathbb{R}^{n\times 3}$ conditioned on the generated fragment structure $\mathbf{x}^{\mathrm{f}}$. We defined a mapping matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$ with $\mathbf{M}_{\mathrm{ik}} = 1$ if the i-th atom belongs to the k-th fragment and 0 otherwise. (Backgrounds and related works of the proposed methods in Appendix A and B, respectively.)

2.1 Fragmentation and 3D fragment structures

We decompose a molecule $G = (\mathcal{V}, \mathcal{E})$ into m non-overlapping fragments $\{S_k\}_{k=1}^m$, where $S_k = (\mathcal{V}_k, \mathcal{E}_k)$ and $\mathcal{V} = \bigcup_{k=1}^m \mathcal{V}_k, \mathcal{E} = \bigcup_{k=1}^m \mathcal{E}_k$ using Principal Subgraph (PS) [27] as illustrated in the step 1 of Fig. 1. Starting from all unique atoms in the fragment vocabulary $\mathcal{S}$, PS iteratively merges neighboring fragments and adds frequent merged fragments to $\mathcal{S}$ until the desired size of $\mathcal{S}$ was reached. The smaller the size of fragment vocabulary, the finer fragments and detailed coarse-grained structures can be obtained.

To generate the initial coordinates of fragments, we utilize RDKit distance geometry [5, 29]. After generating atom coordinates $\hat{\mathbf{x}}^{\mathrm{a}} \sim p_{\mathrm{RDKit}}(\mathbf{x}^{\mathrm{a}})$, we define the fragment coordinates $\mathbf{x}^{\mathrm{f}}$ as averages of their constituent atom coordinates, $\mathbf{M}^\dagger\hat{\mathbf{x}}^{\mathrm{a}}$ where $\mathbf{M}^\dagger$ is a pseudoinverse matrix of $\mathbf{M}$. Since $\hat{\mathbf{x}}^{\mathrm{a}}$ are approximations, the resulting fragment coordinates are also an approximation $\hat{\mathbf{x}}^{\mathrm{f}} \sim p_{\mathrm{RDKit}}(\mathbf{x}^{\mathrm{f}})$. For fragment features $\mathbf{h}^{\mathrm{f}} \in \mathbb{R}^{m\times 3}$, we define a 3-dimensional vector as a frequency histogram of its constituent atom types based on their chemical properties following [36].

2.2 Equivariant blurring diffusion

In this subsection, we elaborate on the design of our diffusion model, *Equivariant Blurring Diffusion* (EBD), drawing inspiration from the principles of the heat equation $\frac{\partial}{\partial t}\mathbf{x}(i,j,t) = \Delta\mathbf{x}(i,j,t)$. $\Delta = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ is the Laplacian operator, and $\mathbf{V}^T$ and $\mathbf{\Lambda}$ are discrete cosine transform and a diagonal matrix whose elements are eigenvalues of Δ, respectively. EBD is designed to generate fine details of conformers $\mathbf{x}^{\mathrm{a}}$, starting from a coarse-grained, approximate structure $\hat{\mathbf{x}}^{\mathrm{f}}$ and a molecular graph G. The overall scheme of EBD is illustrated in the step 2 of Fig. 1.

2.2.1 Forward process and blurring schedule

We define the data corruption of the forward process as a blurring operation that gradually shifts ground truth atom positions $\mathbf{x}_0^{\mathrm{a}} \sim q(\mathbf{x}_0^{\mathrm{a}})$ to their corresponding fragment coordinates:

$$q(\mathbf{x}_t^{\mathrm{a}}|\mathbf{x}_0^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}) = \mathcal{N}(\mathbf{x}_t^{\mathrm{a}}|f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, t), \sigma^2\mathbf{I}), \tag{1}$$

where $f_{\mathbf{B}}$ is a deterministic blurring operator. Consequently, every atom will be positioned according to its fragment coordinates $\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}$ in the prior fragment structure distribution.

When defining $f_{\mathbf{B}}$ for the forward process, we cannot directly adopt the spectral blurring operator $\mathbf{V}\exp(-\mathbf{\Lambda}t)\mathbf{V}^T$ of IHDM [40] for the two reasons: i) For a single molecule, we need to calculate and decompose the fragment graph Laplacian $\{\mathbf{V}_k\mathbf{\Lambda}_k\mathbf{V}_k^T\}_{k=1}^m$ for each fragment $S_k = (\mathcal{V}_k, \mathcal{E}_k)$, unlike a single Laplacian operator per image in IHDM. Given the varying sizes and structures across fragments, it becomes challenging to uniformly adjust the movement of atoms across all fragments using a function of $\{\mathbf{\Lambda}_k\}_{k=1}^m$ in spectral space. ii) As $t \to T$, the ground truth atom coordinates $\mathbf{x}_0^{\mathrm{a}}$ will converge to the ground truth scaffold structure $\mathbf{x}^{\mathrm{f}} = \mathbf{V}\exp(-\mathbf{\Lambda}T)\mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}$ by spectral graph theory [4]. However, there exists a mismatch between the ground truth coordinates $\mathbf{x}^{\mathrm{f}}$ and the approximation coordinates $\hat{\mathbf{x}}^{\mathrm{f}}$ from RDKit in the generative processes. This distributional shift of the fragment structure can potentially harm the performance during the inference.

To circumvent these issues, we transition the space of the blurring operator from spectral domain to spatial domain while retaining the essence of the blurring process. We define $f_{\mathbf{B}}$ as a linear interpolation between $\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}$ and $\mathbf{x}_0^{\mathrm{a}}$ in Euclidean space:

$$f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, t) = (1 - \frac{t}{T})\mathbf{x}_0^{\mathrm{a}} + \frac{t}{T}\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}. \tag{2}$$

As t progresses from 0 to T, the atom coordinates $\mathbf{x}_t^{\mathrm{a}}$ will gradually converge to the fragment structure $\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}$, allowing for uniform adjustment of atom movement. Additionally, we can mitigate the need for excessive eigendecomposition of the fragment graph Laplacian. The example of our blurring schedule on a single fragment is depicted in the step 2 of Fig. 1.

2.2.2 Reverse process and deblurring networks

The aim of the reverse process is to generate fine details at the atom-level from a prior distribution of 3D fragment structure $p(\mathbf{x}_T^{\mathrm{a}}) = \mathcal{N}(\mathbf{x}_T^{\mathrm{a}}|\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}, \delta^2\mathbf{I})$ that is roto-translational invariant to the group. Drawing upon proofs regarding the conditions for an invariant likelihood [26, 55], we develop the deblurring process on the zero center-of-mass subspace using equivariant transition distributions:

$$p_\theta(\mathbf{x}_{t-1}^{\mathrm{a}}|\mathbf{x}_t^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, G) = \mathcal{N}(\mathbf{x}_{t-1}^{\mathrm{a}}|\mu_\theta(\mathbf{x}_t^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, G, t), \delta^2\mathbf{I}), \tag{3}$$

where μ_θ is a parameterized mean function consisting of a deblurring network. To ensure equivariance in the transition distribution, we devise μ_θ inspired by equivariant networks [43]. Our equivariant deblurring network updates invariant features of fragments and atoms $\mathbf{h}^{\mathrm{f}}, \mathbf{h}^{\mathrm{a}}$, and the equivariant coordinates of atoms $\mathbf{x}^{\mathrm{a}}$ by leveraging the hierarchical relationship between atoms and fragments. We consider a complete graph for fragment-level interactions and expand the edge set by incorporating multi-hop and radius neighbors for atom-level interactions. (Details in the Appendix C.)

2.2.3 Training

Following IHDM [40], our loss of previous deblurred state estimation can be defined as:

$$L_{t-1} = \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, t-1) - \rho\big(\mu_\theta(\mathbf{x}_t^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, G, t)\big)\|^2], \tag{4}$$

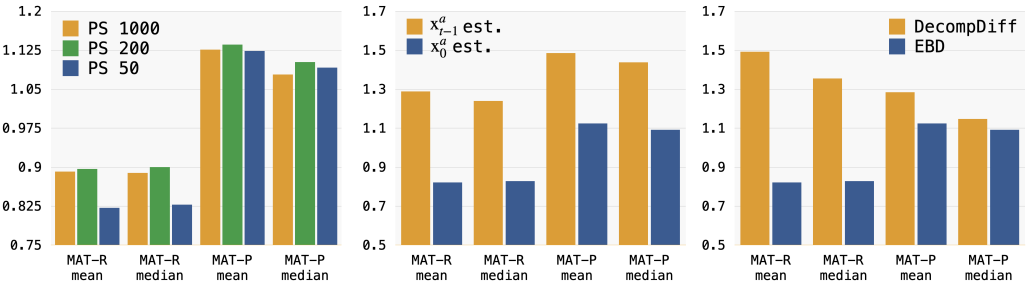
where ρ is the Kabsch algorithm [23] to obtain the optimal rotation matrix for alignment. Through alignment ρ between the prediction from μ_θ and less blurred state $f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, t-1)$ after translating both terms to the zero center-of-mass subspace, the loss function becomes invariant to the SE(3)-transformation of the prediction.

However, we empirically observed that this previous state estimator generates unsatisfactory conformers, similar to the unsatisfactory FID scores observed in image generation of IHDM [40]. We conjectured the reason as the model limited to learn the locally small steps towards the ground truth distribution at each time step [7]. Thus, we reparameterize $\mu_\theta(\mathbf{x}_t^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, G, t)$ as $(1 - \frac{t-1}{T})f_\theta(\mathbf{x}_t^{\mathrm{a}}, G, t) + \frac{t-1}{T}\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}$ to make the deblurring network estimates the ground truth state $\mathbf{x}_0^{\mathrm{a}}$ instead of the previous less blurred state via neural networks f_θ (Derivation in Appendix D.):

$$L_{t-1} = \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}}, \hat{\mathbf{x}}^{\mathrm{f}}, t-1) - \rho\big((1 - \frac{t-1}{T})f_\theta(\mathbf{x}_t^{\mathrm{a}}, G, t) + \frac{t-1}{T}\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}\big)\|^2] \tag{5}$$

$$\approx \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|\mathbf{x}_0^{\mathrm{a}} - \rho\big(f_\theta(\mathbf{x}_t^{\mathrm{a}}, G, t)\big)\|^2]. \tag{6}$$

By loss reparameterization, ρ aligns the prediction to the ground truth state. At time step t of the sampling process, after estimating ground truth $\tilde{\mathbf{x}}_0^{\mathrm{a}}$ from $\mathbf{x}_t^{\mathrm{a}}$, the next state $\mathbf{x}_{t-1}^{\mathrm{a}}$ is computed from a deterministic blurring function $f_{\mathbf{B}}$ using $\tilde{\mathbf{x}}_0^{\mathrm{a}}$.



(a) What level of granularity? (b) What to estimate? (c) How to corrupt?

Figure 2: Ablation studies on the motivation and design choice of EBD. (a) Fragment vocabulary granularity; (b) Target of state estimator; (c) Choice of data corruption processes.

3 Experiments

We use GEOM-Drugs (Drugs) [1] which is drug-like molecules. Drugs comprises 40,000 molecules for the training set and 5,000 molecules for the validation set, with each molecule containing 5 conformers following data split of [44]. For the test set, we selected 200 molecules resulting in 14,324 conformers. To measure the accuracy and diversity of the generated conformer set $\mathcal{C}$, we adopted metrics which measure coverage (COV) and matching (MAT) scores, COV-R, MAT-R, COV-P and MAT-P, proposed by [10]. The metrics are based on root-mean-square deviation, which is a normalized Frobenius norm between two atomic coordinate matrices aligned using the Kabsch algorithm [23]. For each molecule, we generated conformers C that are twice the size of the ground truth conformers C^*. For EBD, we use the $T = 50$, a noise scale of 0.01 for the forward process (σ in Eq. 1) and 0.0125 for the reverse process (δ in Eq. 3) for every experiments.

We compare EBD to existing deep generative models for molecular conformer. The performance of RDKit [29] that was used to generate the fragment structure of our model was measured as a baseline. Besides of RDKit, machine learning models including CVGAE [32], GraphDG [45], CGCF [53], ConfVAE [54], GeoMol [10], ConfGF [44], and GeoDiff [55] were compared to our model. In a case of GeoDiff, we adhered to their settings by configuring the maximum time step T to 5,000. (Implementation details in Appendix E.)

3.1 Ablation studies

We conducted ablation studies to validate our model design, encompassing the size of the fragment vocabulary, the reparameterization of loss, and the blurring data corruption. For each ablation study, we calculated the mean and median of matching scores MAT-R and MAT-P. Note that lower values of MAT-R and MAT-P indicate better results.

Effects of fragment granularity. We assessed the performance variation as fragment structure became more detailed and informative by measuring the generation performances across different fragment vocabulary sizes $|\mathcal{S}| \in \{50, 200, 1000\}$. Since PS [27], the fragmentation method we used, initializes the vocabulary from unique single atoms, reducing the size $|\mathcal{S}|$ results in obtaining finer fragments $\hat{\mathbf{x}}^f$. We reported the statistics of the average number of fragments per graph ($\#\text{frags}/G$) and atoms per fragment ($\#\text{atoms}/\text{frag}$) in Table 1 and the generation results in Fig. 2 (a). Thanks to the increased level of detail in fragments, $|\mathcal{S}| = 50$ can obtain better performance compared to other vocabulary sizes. This is because more specific fragment structures decrease the amount of atomic-level detail that needs to be generated. From this observation, we use $|\mathcal{S}| = 50$ in all subsequent experiments.

Table 1: Statistics of fragment vocabulary $\mathcal{S}$ in Drugs.

$\lvert\mathcal{S}\rvert$	$\#\text{frags}/G$	$\#\text{atoms}/\text{frag}$
50	11.77	4.02
200	7.60	6.34
1000	5.26	9.25

Effects of loss reparameterization. We presented the performance comparison between the less blurred previous state estimator in Eq. (4) and ground truth estimator in Eq. (6) after loss reprarameterization in Fig. 2 (b). From the previous state estimator, we acquired degenerated conformers with relatively high matching scores, which align with low FID score of IHDM [40] in image generation. On the other hand, we observed distinct advantages in introducing the ground truth estimator across

Table 2: Geometric evaluation on Drugs ($\delta = 1.25$Å).

Models	COV-R (%) ↑ Mean	COV-R (%) ↑ Med	MAT-R(Å) ↓ Mean	MAT-R(Å) ↓ Med	COV-P (%) ↑ Mean	COV-P (%) ↑ Med	MAT-P (Å) ↓ Mean	MAT-P (Å) ↓ Med
RDKit	45.74	31.75	1.5376	1.4004	54.78	59.48	1.3341	1.1996
CVGAE	0.00	0.00	3.0702	2.9937	-	-	-	-
GraphDG	8.27	0.00	1.9722	1.9845	2.08	0.00	2.4340	2.4100
CGCF	53.96	57.06	1.2487	1.2247	21.68	13.72	1.8571	1.8066
ConfVAE	55.20	59.43	1.2380	1.1417	22.96	14.05	1.8287	1.8159
GeoMol	67.16	71.71	1.0875	1.0586	-	-	-	-
ConfGF	62.15	70.93	1.1629	1.1596	23.42	15.52	1.7219	1.6863
GeoDiff	89.40	96.86	0.8571	0.8495	61.28	65.00	1.1642	1.1272
EBD	**92.60**	**98.73**	**0.8216**	**0.8279**	**66.24**	**68.39**	**1.1237**	**1.0916**

all metrics. We speculate that the ground truth estimator facilitates the diffusion model in learning beyond locally blurring distributions towards the target distribution.

Effects of data corruptions. We provide the same initial fragment structures $\hat{\mathbf{x}}^f$ to both EBD and DecompDiff [12] so that the data corruption method becomes the primary distinction to examine. DecompDiff denoises multiple prior distributions, where each mean corresponds to the coordinates of each fragment $\hat{\mathbf{x}}^f$. The generation results and sampling trajectories are compared between the two models ($T = 50$ for both) in Fig. 2 (c). At first, we observed that the conformers generated from DecompDiff exhibit lower diversity scores compared to EBD. This is because the results of DecompDiff tend to adhere closely to the approximate fragment structure $\hat{\mathbf{x}}^f$, whereas EBD attempts to transition towards the ground truth fragment structure $\mathbf{x}^f$. We speculate that our blurring schedule, which entails a linear interpolation between $\mathbf{M}\hat{\mathbf{x}}^f$ and $\mathbf{x}_0^a$, facilitates the learning process for the diffusion model compared to a stochastic trajectory between prior and target distributions. As empirical evidence, we observed that DecompDiff primarily focuses on denoising the fragment structure throughout most of the sampling process in Fig. 3. On the other hand, EBD focuses on the entirety of the sampling process to generate fine details, resulting in better quality.

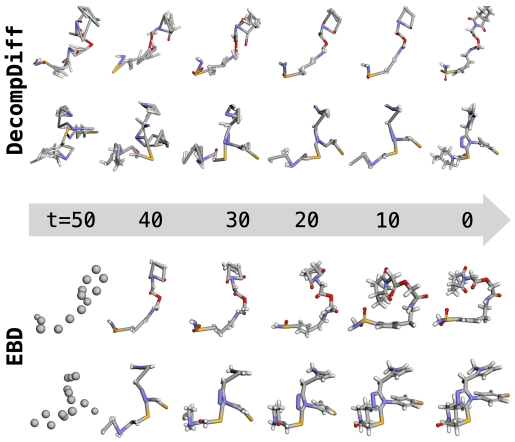


Figure 3: Sampling processes of two conformers depending on data corruptions.

3.2 Geometric evaluation

We compared our hierarchical framework to the baseline RDKit and machine learning models for molecular conformer generation on Drugs, and the results are reported in Table 2. EBD achieves superior performance across all metrics by generating diverse and accurate molecular conformers. In comparison to RDKit, which was used to generate fragment structures $\hat{\mathbf{x}}^f$, EBD achieved a significant improvement in the generation of diverse fine atomic details, as evidenced by higher COV-R and MAT-R scores. We also observed that, due to the informative fragment structure prior distribution and the proposed blurring schedule, EBD produces more diverse and higher-quality conformers even with 100 times fewer T compared to GeoDiff. (Further experiment results in Appendix F and G.)

4 Conclusion

We introduced a novel hierarchical generative model for molecular conformers via Equivariant Blurring Diffusion (EBD), a diffusion model designed for coarse-to-fine generative scheme. After generating the initial distribution of fragment coordinates from a cheminformatics tool, EBD generated fine atomic details from coarse-grained structures through equivariant networks. We also proposed a simple and effective linear blurring scheduler and ground truth state estimator to enhance the model's ability to produce diverse and accurate conformers. Through extensive analysis of the proposed model and comparison between denoising diffusion models, we substantiated the validity of the model design.

References

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

[2] Bansal, Arpit, Borgnia, Eitan, Chu, Hong-Min, Li, Jie, Kazemi, Hamid, Huang, Furong, Goldblum, Micah, Geiping, Jonas, and Goldstein, Tom. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36, 2023.

[3] Campbell, Andrew, Harvey, William, Weilbach, Christian, De Bortoli, Valentin, Rainforth, Thomas, and Doucet, Arnaud. Trans-dimensional generative modeling via jump diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

[4] Chung, Fan RK. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[5] Crippen, Gordon M, Havel, Timothy F, et al. *Distance geometry and molecular conformation*, volume 74. Research Studies Press Taunton, 1988.

[6] Crouse, David F. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.

[7] Daras, Giannis, Delbracio, Mauricio, Talebi, Hossein, Dimakis, Alexandros, and Milanfar, Peyman. Soft diffusion: Score matching with general corruptions. *Transactions on Machine Learning Research*, 2023.

[8] Daras, Giannis, Shah, Kulin, Dagan, Yuval, Gollakota, Aravind, Dimakis, Alex, and Klivans, Adam. Ambient diffusion: Learning clean distributions from corrupted data. *Advances in Neural Information Processing Systems*, 36, 2023.

[9] Degen, Jorg, Wegscheid-Gerlach, Christof, Zaliani, Andrea, and Rarey, Matthias. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503, 2008.

[10] Ganea, Octavian, Pattanaik, Lagnajit, Coley, Connor, Barzilay, Regina, Jensen, Klavs, Green, William, and Jaakkola, Tommi. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems*, 34, 2021.

[11] Geng, Zijie, Xie, Shufang, Xia, Yingce, Wu, Lijun, Qin, Tao, Wang, Jie, Zhang, Yongdong, Wu, Feng, and Liu, Tie-Yan. De novo molecular generation via connection-aware motif mining. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Guan, Jiaqi, Zhou, Xiangxin, Yang, Yuwei, Bao, Yu, Peng, Jian, Ma, Jianzhu, Liu, Qiang, Wang, Liang, and Gu, Quanquan. Decompdiff: Diffusion models with decomposed priors for structure-based drug design. In *International Conference on Machine Learning*, pp. 11827–11846. PMLR, 2023.

[13] Heusel, Martin, Ramsauer, Hubert, Unterthiner, Thomas, Nessler, Bernhard, and Hochreiter, Sepp. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[14] Ho, Jonathan, Jain, Ajay, and Abbeel, Pieter. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 2020.

[15] Ho, Jonathan, Saharia, Chitwan, Chan, William, Fleet, David J, Norouzi, Mohammad, and Salimans, Tim. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

[16] Hono, Yukiya, Tsuboi, Kazuna, Sawada, Kei, Hashimoto, Kei, Oura, Keiichiro, Nankaku, Yoshihiko, and Tokuda, Keiichi. Hierarchical multi-grained generative model for expressive speech synthesis. *Interspeech 2020*, 2020.

[17] Hoogeboom, Emiel and Salimans, Tim. Blurring diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.

[18] Hoogeboom, Emiel, Satorras, Vıctor Garcia, Vignac, Clément, and Welling, Max. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

[19] Hsu, Wei-Ning, Zhang, Yu, Weiss, Ron, Zen, Heiga, Wu, Yonghui, Cao, Yuan, and Wang, Yuxuan. Hierarchical generative modeling for controllable speech synthesis. In *International Conference on Learning Representations*, 2019.

[20] Jin, Wengong, Barzilay, Regina, and Jaakkola, Tommi. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, pp. 2323–2332. PMLR, 2018.

[21] Jin, Wengong, Barzilay, Regina, and Jaakkola, Tommi. Hierarchical generation of molecular graphs using structural motifs. In *International Conference on Machine Learning*, pp. 4839–4848. PMLR, 2020.

[22] Jing, Bowen, Corso, Gabriele, Chang, Jeffrey, Barzilay, Regina, and Jaakkola, Tommi. Torsional diffusion for molecular conformer generation. *Advances in Neural Information Processing Systems*, 35, 2022.

[23] Kabsch, Wolfgang. 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, 1976.

[24] Karras, Tero, Aittala, Miika, Aila, Timo, and Laine, Samuli. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35, 2022.

[25] Kingma, Diederik, Salimans, Tim, Poole, Ben, and Ho, Jonathan. Variational diffusion models. *Advances in Neural Information Processing Systems*, 34, 2021.

[26] Köhler, Jonas, Klein, Leon, and Noé, Frank. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International Conference on Machine Learning*, pp. 5361–5370. PMLR, 2020.

[27] Kong, Xiangzhe, Huang, Wenbing, Tan, Zhixing, and Liu, Yang. Molecule generation by principal subgraph mining and assembling. *Advances in Neural Information Processing Systems*, 35, 2022.

[28] Krizhevsky, Alex, Hinton, Geoffrey, et al. Learning multiple layers of features from tiny images. 2009.

[29] Landrum, Greg et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

[30] Lewell, Xiao Qing, Judd, Duncan B, Watson, Stephen P, and Hann, Michael M. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of chemical information and computer sciences*, 38(3):511–522, 1998.

[31] Loshchilov, Ilya and Hutter, Frank. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[32] Mansimov, Elman, Mahmood, Omar, Kang, Seokho, and Cho, Kyunghyun. Molecular geometry prediction using a deep generative graph neural network. *Scientific reports*, 9(1):20381, 2019.

[33] Menick, Jacob and Kalchbrenner, Nal. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In *International Conference on Learning Representations*, 2019.

[34] Nichol, Alexander Quinn and Dhariwal, Prafulla. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

[35] Paszke, Adam, Gross, Sam, Chintala, Soumith, Chanan, Gregory, Yang, Edward, DeVito, Zachary, Lin, Zeming, Desmaison, Alban, Antiga, Luca, and Lerer, Adam. Automatic differentiation in pytorch. 2017.

[36] Qiang, Bo, Song, Yuxuan, Xu, Minkai, Gong, Jingjing, Gao, Bowen, Zhou, Hao, Ma, Wei-Ying, and Lan, Yanyan. Coarse-to-fine: a hierarchical diffusion model for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 28277–28299. PMLR, 2023.

[37] Ramakrishnan, Raghunathan, Dral, Pavlo O, Rupp, Matthias, and Von Lilienfeld, O Anatole. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

[38] Razavi, Ali, Van den Oord, Aaron, and Vinyals, Oriol. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] Reidenbach, Danny and Krishnapriyan, Aditi. Coarsenconf: Equivariant coarsening with aggregated attention for molecular conformer generation. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[40] Rissanen, Severi, Heinonen, Markus, and Solin, Arno. Generative modelling with inverse heat dissipation. In *The Eleventh International Conference on Learning Representations*, 2023.

[41] Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Björn. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[42] Saharia, Chitwan, Chan, William, Saxena, Saurabh, Li, Lala, Whang, Jay, Denton, Emily L, Ghasemipour, Kamyar, Gontijo Lopes, Raphael, Karagol Ayan, Burcu, Salimans, Tim, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35, 2022.

[43] Satorras, Vıctor Garcia, Hoogeboom, Emiel, and Welling, Max. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*, pp. 9323–9332. PMLR, 2021.

[44] Shi, Chence, Luo, Shitong, Xu, Minkai, and Tang, Jian. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, pp. 9558–9568. PMLR, 2021.

[45] Simm, Gregor and Hernandez-Lobato, Jose Miguel. A generative model for molecular distance geometry. In *International Conference on Machine Learning*, pp. 8949–8958. PMLR, 2020.

[46] Sohl-Dickstein, Jascha, Weiss, Eric, Maheswaranathan, Niru, and Ganguli, Surya. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

[47] Song, Jiaming, Meng, Chenlin, and Ermon, Stefano. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

[48] Song, Yang and Ermon, Stefano. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] Song, Yang, Sohl-Dickstein, Jascha, Kingma, Diederik P, Kumar, Abhishek, Ermon, Stefano, and Poole, Ben. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[50] Vahdat, Arash, Kreis, Karsten, and Kautz, Jan. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34, 2021.

[51] Vahdat, Arash, Williams, Francis, Gojcic, Zan, Litany, Or, Fidler, Sanja, Kreis, Karsten, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022.

[52] Wang, Wujie, Xu, Minkai, Cai, Chen, Miller, Benjamin K, Smidt, Tess, Wang, Yusu, Tang, Jian, and Gomez-Bombarelli, Rafael. Generative coarse-graining of molecular conformations. In *International Conference on Machine Learning*, pp. 23213–23236. PMLR, 2022.

[53] Xu, Minkai, Luo, Shitong, Bengio, Yoshua, Peng, Jian, and Tang, Jian. Learning neural generative dynamics for molecular conformation generation. In *International Conference on Learning Representations*, 2021.

[54] Xu, Minkai, Wang, Wujie, Luo, Shitong, Shi, Chence, Bengio, Yoshua, Gomez-Bombarelli, Rafael, and Tang, Jian. An end-to-end framework for molecular conformation generation via bilevel programming. In *International Conference on Machine Learning*, pp. 11537–11547. PMLR, 2021.

[55] Xu, Minkai, Yu, Lantao, Song, Yang, Shi, Chence, Ermon, Stefano, and Tang, Jian. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

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

[57] Yang, Soojung and Gómez-Bombarelli, Rafael. Chemically transferable generative backmapping of coarse-grained proteins. In *International Conference on Machine Learning*, pp. 39277–39298, 2023.

[58] Zhu, Jinhua, Xia, Yingce, Liu, Chang, Wu, Lijun, Xie, Shufang, Wang, Yusong, Wang, Tong, Qin, Tao, Zhou, Wengang, Li, Houqiang, et al. Direct molecular conformation generation. *Transactions on Machine Learning Research*, 2022.

A Backgrounds

A.1 Blurring diffusion

The denoising diffusion models [46, 47, 49, 14, 25], which corrupt data by adding random noise and generate data through denoising, have significantly advanced across diverse domains [51, 50, 42]. Recently, a few works [2, 40, 7, 17] have introduced data corruption into the design space of diffusion models [24], going beyond random noise corruption in the vision domain.

Inverse Heat Dissipation Model (IHDM) [40] proposed a coarse-to-fine generation in the pixel space. Their forward process follows a partial differential equation of heat dissipation on grids:

$$\frac{\partial}{\partial t}\mathbf{x}(i,j,t) = \Delta\mathbf{x}(i,j,t), \tag{A.1}$$

where $\mathbf{x}$ represents the data on the grid and Δ is the Laplacian operator. IHDM derived the solution of this equation at time step t, $\mathbf{x}_t$, using eigendecomposition of Δ as:

$$\mathbf{x}_t = \mathbf{B}_t\mathbf{x}_0 = \mathbf{V}\exp(-\mathbf{\Lambda}t)\mathbf{V}^T\mathbf{x}_0, \tag{A.2}$$

where $\mathbf{V}^T$ and $\mathbf{\Lambda}$ are discrete cosine transform and a diagonal matrix whose elements are eigenvalues of Δ, respectively. As $t \to T$, the eigenbasis of eigenvalue 0 only remains and this leads to the convergence of pixel intensities to their average value. Based upon this blurring process, IHDM defined a forward process as:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\mathbf{B}_t\mathbf{x}_0, \sigma^2\mathbf{I}), \tag{A.3}$$

which means that the state at t is equal to the data blurred until t with small amount of noise. Note that the function of data corruption $\mathbf{B}_t$ was defined at a spectral space of eigenvalues $\mathbf{\Lambda}$. Then, the reverse generative process was defined to deblur each state:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}|\mu_\theta(\mathbf{x}_t, t), \delta^2\mathbf{I}), \tag{A.4}$$

where the mean at $t-1$ is the result of a deblurring network μ_θ and δ is the small amount of standard deviation for noise. As t approaches 0, μ_θ gradually restores fine details from coarse-grained information about pixel intensities by effectively deblurring state values. The loss was defined to minimize the distance between the result of deblurring network and less blurred state at randomly sampled t as:

$$L_{t-1} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t}[\|\mathbf{B}_{t-1}\mathbf{x}_0 - \mu_\theta(\mathbf{x}_t, t)\|^2]. \tag{A.5}$$

IHDM was evaluated on image generation task using FID score [13], but its performance lagged behind that of denoising diffusion models. For instance, IHDM achieves an FID score of 18.96 while DDPM [14] have 3.17 on CIFAR-10 [28].

A.2 Equivariance

In this work, we consider the SE(3) group to address the roto-translational equivariance of molecular conformers [26, 18, 55]. A function f is equivariant to a group $\mathcal{G}$ if $T_g(f(\mathbf{x})) = f(S_g(\mathbf{x}))$ holds for all $g \in \mathcal{G}$, where T_g, S_g are transformations of the group element g. In our coarse-to-fine generative framework, the invariant prior distribution of coarse-grained structure represents the coordinates of fragments. Therefore, the design of the transition distribution and the loss function in our diffusion model need to ensure that the generated likelihood is invariant, so that the generated conformers are not affected by rotation or translation.

B Related work

B.1 Multiscale generation

A multiscale design of generative models is evident across multiple domains, including image generation [33, 38, 15, 40] and speech synthesis [19, 16], aimed at enhancing the interpretability and quality of samples derived from coarse-grained information. In the field of computational biology, recent studies on molecular graph generation [20, 21, 11], backmapping of protein structure [57] and conformer generation [52] conditioned on the given ground truth coarse-grained information have reported the effectiveness of the multiscale design. In recent unconditional conformer generation [36], a denoising diffusion model [18] was exclusively used in the fragment structure generation step and not designed for coarse-to-fine generation.

B.2 Data corruption in diffusion models

The choice of data corruption can be considered a crucial aspect of the design space of diffusion models [24], depending on the characteristics of the data domain and the specific problem definition. Recently, several studies on diffusion models have revealed that the choice of data corruption can be extended beyond random noise [46, 14, 47, 34, 41] to methods such as masking [2, 7, 8], blurring [40, 2, 7, 17], and varying data dimension [3]. We designed the data corruption process as a blurring operation in Euclidean space, transitioning from atom-level fine details to fragment-level coarse structures. This approach is more effective for multiscale frameworks compared to random noise, which corrupts both fragment and atom geometries.

C Deblurring network architectures

In SE(3)-equivariant deblurring networks, there are update functions of SE(3)-invariant fragment and atom features $\mathbf{h}^{\mathrm{f}}$, $\mathbf{h}^{\mathrm{a}}$, as well as an update function of SE(3)-equivariant atom coordinates $\mathbf{x}^{\mathrm{a}}$ motivated from equivariant graph neural networks [43]. For the fragments, we constructed a complete graph to account for dense interactions among them. In the case of atoms, we expanded the neighbor set of each atom by including multi-hop neighbors derived from the powers of the adjacency matrix and a radius graph, which includes atoms within a specified cutoff distance. The benefits of dense interactions for accurate conformers estimation have been confirmed in several studies [45, 55, 18].

The architecture of the SE(3)-invariant message passing and feature update functions at the fragment- and atom-level is as follows:

$$\mathbf{m}^{\mathrm{f}}_{\mathtt{ij}} = \phi^{\mathrm{f}}_m(\mathbf{h}^{\mathrm{f},l}_{\mathtt{i}}, \mathbf{h}^{\mathrm{f},l}_{\mathtt{j}}, \|\mathbf{x}^{\mathrm{f}}_{\mathtt{i}} - \mathbf{x}^{\mathrm{f}}_{\mathtt{j}}\|), \qquad \mathbf{h}^{\mathrm{f},l+1}_{\mathtt{i}} = \phi^{\mathrm{f}}_h(\mathbf{h}^{\mathrm{f},l}_{\mathtt{i}}, \sum\nolimits_{\mathtt{j}\in N(\mathbf{x}^{\mathrm{f}}_{\mathtt{i}})} \mathbf{m}^{\mathrm{f}}_{\mathtt{ij}}, \mathbf{h}^{\mathrm{a},l}), \tag{A.6}$$

$$\mathbf{m}^{\mathrm{a}}_{\mathtt{ij}} = \phi^{\mathrm{a}}_m(\mathbf{h}^{\mathrm{a},l}_{\mathtt{i}}, \mathbf{h}^{\mathrm{a},l}_{\mathtt{j}}, \|\mathbf{x}^{\mathrm{a},l}_{\mathtt{i}} - \mathbf{x}^{\mathrm{a},l}_{\mathtt{j}}\|, e^{\mathrm{a}}_{\mathtt{ij}}), \quad \mathbf{h}^{\mathrm{a},l+1}_{\mathtt{i}} = \phi^{\mathrm{a}}_h(\mathbf{h}^{\mathrm{a},l}_{\mathtt{i}}, \sum\nolimits_{\mathtt{j}\in N(\mathbf{x}^{\mathrm{a}}_{\mathtt{i}})} \mathbf{m}^{\mathrm{a}}_{\mathtt{ij}}, \mathbf{h}^{\mathrm{f},l+1}), \tag{A.7}$$

where $\mathbf{m}_{\mathtt{ij}} \in \mathbb{R}^d$ is the message for each interactions, and $\mathbf{h} \in \mathbb{R}^d$ is the feature vector from the aggregated messages and features from different hierarchy level. For every invariant update functions $\phi^{\mathrm{f}}_m, \phi^{\mathrm{f}}_h, \phi^{\mathrm{a}}_m, \phi^{\mathrm{a}}_h$, we used multilayer perceptrons. For initial features $\mathbf{h}^{\mathrm{f},0}_{\mathtt{i}}$ of fragments, we defined a 3-dimensional vector as a frequency histogram of its constituent atom types based on their chemical properties, including hydrophobicity, hydrogen bond center, and negative charge center following [36]. The detailed definition of the initial fragment features is in Table 3. For initial atom features $\mathbf{h}^{\mathrm{a},0}_{\mathtt{i}} \in \mathbb{R}^d$ and bond features $e^{\mathrm{a}}_{\mathtt{ij}} \in \mathbb{R}^d$, we used embeddings from atom types and bond types, respectively.

Table 3: Initial fragment feature based on chemical properties.

Properties	Details	Types
Hydrophobicity	Frequency of C element	Integer
Hydrogen bond center	Frequency of O, N, S, P elements	Integer
Negative charge center	Frequency of F, Cl, Br, I elements	Integer

For the i-th atom $\mathbf{x}^{\mathrm{a}}_{\mathtt{i}}$ belongs to the k-th fragment $\mathbf{x}^{\mathrm{f}}_{\mathtt{k}}$, the architecture of the equivariant atom coordinate update function is as follows:

$$\begin{aligned}\mathbf{x}^{\mathrm{a},l+1}_{\mathtt{i}} = \mathbf{x}^{\mathrm{a},l}_{\mathtt{i}} &+ \sum_{\mathtt{j}\in N(\mathbf{x}^{\mathrm{a}}_{\mathtt{i}})} \frac{\mathbf{x}^{\mathrm{a},l}_{\mathtt{i}} - \mathbf{x}^{\mathrm{a},l}_{\mathtt{j}}}{d^{\mathrm{a},l}_{\mathtt{ij}} + 1} \phi^{\mathrm{a}}_x(\mathbf{h}^{\mathrm{a},l+1}_{\mathtt{i}}, \mathbf{h}^{\mathrm{a},l+1}_{\mathtt{j}}, \mathbf{m}^{\mathrm{a}}_{\mathtt{ij}}, e^{\mathrm{a}}_{\mathtt{ij}}) \\ &+ \frac{\mathbf{x}^{\mathrm{a},l}_{\mathtt{i}} - \mathbf{x}^{\mathrm{f}}_{\mathtt{k}}}{\|\mathbf{x}^{\mathrm{a},l}_{\mathtt{i}} - \mathbf{x}^{\mathrm{f}}_{\mathtt{k}}\| + 1} \phi^{\mathrm{f}}_x(\mathbf{h}^{\mathrm{a},l+1}_{\mathtt{i}}, \mathbf{h}^{\mathrm{f},l+1}_{\mathtt{k}}, \|\mathbf{x}^{\mathrm{a},l}_{\mathtt{i}} - \mathbf{x}^{\mathrm{f}}_{\mathtt{k}}\|),\end{aligned} \tag{A.8}$$

where $\mathbf{x}^{\mathrm{f}}_{\mathtt{k}}$ is the k-th row of $\mathbf{M}^{\dagger}\mathbf{x}^{\mathrm{a}}_t$, and $d^{\mathrm{a},l}_{\mathtt{ij}} = \|\mathbf{x}^{\mathrm{a},l}_{\mathtt{i}} - \mathbf{x}^{\mathrm{a},l}_{\mathtt{j}}\|$ are inter-atomic distances. For every equivariant update functions $\phi^{\mathrm{a}}_x, \phi^{\mathrm{f}}_x$, we used multilayer perceptrons. For three terms in right-hand side of Eq. A.8, the first term is the coordinate from the previous layer, the second term is an equivariant update function that accounts for atom-level interactions, and the third term is an equivariant update function that considers the deviation of the current atom coordinate from its respective fragment's coordinate.

D Derivation of loss function

In this section, we explain the derivation of the loss function for the ground truth state estimator from the previous state estimator. The loss function of previous state estimation is defined as:

$$L_{t-1} = \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}},t-1) - \rho\big(\mu_\theta(\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}},G,t)\big)\|^2], \tag{A.9}$$

where ρ is the Kabsch algorithm [23] to obtain the optimal rotation matrix for alignment. Through alignment ρ of the prediction from μ_θ to the less blurred state $f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}},t-1)$ after translating both terms to the zero center-of-mass subspace, the loss function becomes invariant to the translation and rotation of the prediction.

However, this previous state estimator generates unsatisfactory conformers as empirically observed in Sec. 3.1. We conjectured the reason as the model limited to learn the locally small steps towards the ground truth distribution at each time step [7]. Thus, we reparameterize $\mu_\theta(\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}},G,t)$ as $(1-\frac{t-1}{T})f_\theta(\mathbf{x}_t^{\mathrm{a}},G,t) + \frac{t-1}{T}\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}$ to make the deblurring network estimates the ground truth state $\mathbf{x}_0^{\mathrm{a}}$ instead of the previous less blurred state via neural networks f_θ. We first start with the non-invariant previous state estimation, which is without the alignment ρ:

$$\begin{aligned}
L_{t-1} &= \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}},t-1) - \mu_\theta(\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}},G,t)\|^2], && \text{(A.10)}\\
&= \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|f_{\mathbf{B}}(\mathbf{x}_0^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}},t-1) - (1-\tfrac{t-1}{T})f_\theta(\mathbf{x}_t^{\mathrm{a}},G,t) - \tfrac{t-1}{T}\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}\|^2] && \text{(A.11)}\\
&= \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|(1-\tfrac{t-1}{T})\mathbf{x}_0^{\mathrm{a}} + \tfrac{t-1}{T}\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}} - (1-\tfrac{t-1}{T})f_\theta(\mathbf{x}_t^{\mathrm{a}},G,t) - \tfrac{t-1}{T}\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}}\|^2] && \text{(A.12)}\\
&= \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|(1-\tfrac{t-1}{T})(\mathbf{x}_0^{\mathrm{a}} - f_\theta(\mathbf{x}_t^{\mathrm{a}},G,t)) + \tfrac{t-1}{T}(\mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}} - \mathbf{M}\hat{\mathbf{x}}^{\mathrm{f}})\|^2] && \text{(A.13)}\\
&= \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}(1-\tfrac{t-1}{T})^2[\|\mathbf{x}_0^{\mathrm{a}} - f_\theta(\mathbf{x}_t^{\mathrm{a}},G,t)\|^2] && \text{(A.14)}\\
&\approx \mathbb{E}_{t,\mathbf{x}_0^{\mathrm{a}},\mathbf{x}_t^{\mathrm{a}},\hat{\mathbf{x}}^{\mathrm{f}}}[\|\mathbf{x}_0^{\mathrm{a}} - \rho\big(f_\theta(\mathbf{x}_t^{\mathrm{a}},G,t)\big)\|^2]. && \text{(A.15)}
\end{aligned}$$

In the last stage from Eq. A.14 to Eq. A.15, we simplified the loss function by discarding the time-dependent weight as [14]. Finally, we make the loss function for ground truth estimation invariant by aligning the prediction from f_θ to the ground truth state using Kabsch alignment ρ [23].

E Implementation details

E.1 Datasets

We used GEOM-QM9 (QM9) [37] and GEOM-Drugs (Drugs) [1] for analysis and comparison between molecular conformer generation models. Each dataset comprises 40,000 molecules for the training set and 5,000 molecules for the validation set, with each molecule containing 5 conformers following data split of [44]. We obtained the raw data, the pre-processed data and the data split at `https://github.com/DeepGraphLearning/ConfGF`. For the test set, we selected 200 molecules for each dataset, resulting in 22,408 and 14,324 conformers existing in QM9 and Drugs, respectively.

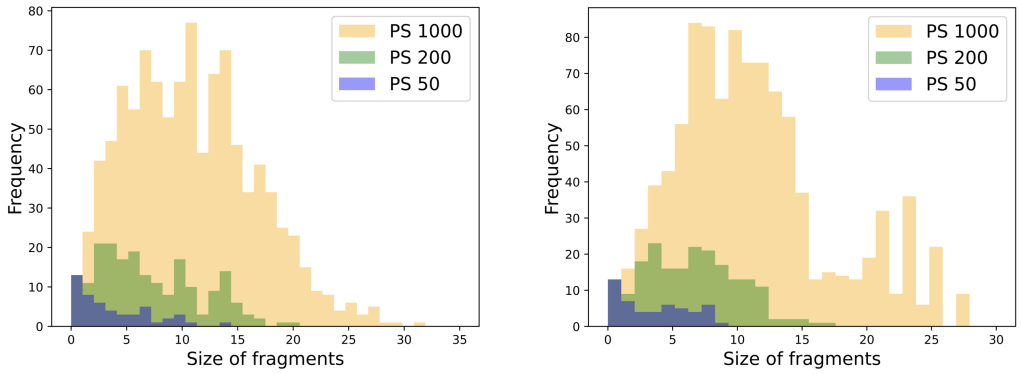
For fragmentation of the molecular graphs $G = (\mathcal{V},\mathcal{E})$ in Drugs and QM9, we used Principal Subgraph (PS) [27] (`https://github.com/THUNLP-MT/PS-VAE`) which can construct a fragment vocabulary $\mathcal{S}$ whose elements are the largest and frequent repetitive subgraphs of molecules. Starting from all unique atoms in $\mathcal{S}$ at initial stage, PS iteratively merges neighboring fragments. The most frequent fragment among the newly merged fragments was added to the vocabulary at each iteration, and this operation was repeated until the desired size of the vocabulary was reached. Thus, the smaller the fragment vocabulary, the finer fragments can be obtained. One of the advantages of PS compared to existing fragmentation methods such as RECAP [30], BRICS [9], junction tree decomposition [20] is the ability to control the vocabulary size, allowing us to observe how performance varies with fragment granularity. We constructed $\mathcal{S}$ for each dataset with three fragment vocabulary sizes $|\mathcal{S}| \in \{50, 200, 1000\}$. The average numbers of fragments per graph ($\#\mathrm{frags}/G$) and atoms per fragment ($\#\mathrm{atoms}/\mathrm{frag}$) of Drugs and QM9 were reported in Table 4.

Additionally, the frequency depends on the size of fragments (number of constituent atoms) in Drugs and QM9 was reported in Fig. 4. For each $|\mathcal{S}|$, the frequency distribution across fragment sizes is smooth and not biased toward certain sizes.

In the training and validation sets of the Drugs and QM9 datasets, there are 5 different ground truth conformers for each molecule. Thus, we generated 5 different conformers from RDKit to compute the

Table 4: Statistics of fragment vocabulary $\mathcal{S}$.

$\lvert\mathcal{S}\rvert$	Drugs		QM9	
	#frags/G	#atoms/frag	#frags/G	#atoms/frag
50	11.77	4.02	5.17	3.91
200	7.60	6.34	3.70	5.45
1000	5.26	9.25	2.91	6.98



(a) Drugs (b) QM9

Figure 4: Frequency depends on the size of fragments in the fragment vocabulary of Drugs and QM9.

fragment coordinates $\hat{\mathbf{x}}^{\mathrm{f}}$ for each molecule in train and validation sets. Following [22], we computed the optimal matching between 5 RDKit generated conformers and 5 different ground truth conformers for a single molecule. After computing the cost matrix whose the (i, j)-th element means RMSD between the i-th RDKit generated conformers and the j-th ground truth conformer, we assigned optimal RDKit conformer to each ground truth conformer using linear sum assignment problem [6]. After finding the optimal matching, we aligned each ground truth conformer to its assigned RDKit conformer using the Kabsch algorithm [23]. The aligned ground truth conformers were then used in the blurring schedule (Eq. 2) and loss function (Eq. 6) of the training process.

E.2 Evalutaion metrics

To measure the accuracy and diversity of the generated conformer set $\mathcal{C}$, we adopted metrics which measure coverage (COV) and matching (MAT) scores, COV-R, MAT-R, COV-P and MAT-P, proposed by [10]. The metrics are based on root-mean-square deviation, which is a normalized Frobenius norm between two atomic coordinate matrices aligned using the Kabsch algorithm [23]. Given the ground truth conformer set $\mathcal{C}^*$ and the generated sample set $\mathcal{C}$, four metrics that follow precision and recall are defined as:

$$\text{COV-R (Recall)} = \frac{1}{|\mathcal{C}^*|}|\{C^* \in \mathcal{C}^* | \text{RMSD}(C^*, C) \leq \delta, C \in \mathcal{C}\}|, \tag{A.16}$$

$$\text{MAT-R (Recall)} = \frac{1}{|\mathcal{C}^*|} \sum_{C^* \in \mathcal{C}^*} \min_{C \in \mathcal{C}} \text{RMSD}(C^*, C), \tag{A.17}$$

where COV and MAT are coverage metric and matching metric [53], respectively. COV quantifies the proportion of one set covered by another, with "covered" indicating RMSD values are within a threshold δ. MAT measures the average of RMSD values of one conformer set with its closest conformer in another set. If $\mathcal{C}$ and $\mathcal{C}^*$ are exchanged in Eqs. (A.16, A.17), then metrics become COV-P (Precision) and MAT-P (Precision). The recall metric is focused on the diversity, while the precision metric measures the quality. The threshold δ is set to 0.5Å for QM9 and 1.25Å for Drugs.

E.3 Training and time

The processes for fragmentation of molecular graphs and obtaining fragment coordinates from RDKit do not harm the efficiency of our framework, as they can be completed before training our diffusion model. To generate 5 distinct fragment coordinates $\hat{\mathbf{x}}^{\mathrm{f}}$ for each of the $45,000$ molecules in the training

and validation set of the GEOM-Drug benchmark [1], it took 38 hours, averaging 3.04 seconds per molecule.

We used a single NVIDIA A100 GPU for every training and generation tasks. For training, we used a learning rate 10^{-4} with the AdamW optimizer [31]. The training time for both Drugs and QM9 was required around 3.8 days. For sampling, Drugs required 145 minutes for 14,324 conformers in 200 molecules, and QM9 required 71 minutes for 22,408 conformers in 200 molecules. We reported hyperparameters of EBD training including the maximum time step T, number of layers ($\# \, l$) and number of features ($\# \, d$) in the deblurring networks, number of multi-hops ($\#$ of hops) and cutoff value for the expansion of atom interactions, batch size, and number of iteration in Table 5.

Table 5: Hyperparameters of EBD.

Dataset	T	$\# \, l$	$\# \, d$	$\#$ of hops	cutoff	batch size	training iter.
Drugs	50	6	128	3	10 Å	32	650k
QM9	50	6	128	3	10 Å	64	650k

E.4 Performance of compared methods

For the results of compared methods in geometric evaluation of Drugs (Table 2) and QM9 (Table 6), COV-R and MAT-R scores of CVGAE [32], GraphDG [45], CGCF [53], and ConfGF [44] were borrowed from [44]. The performance of GeoMol and ConfVAE were borrowed from [58] and [55], respectively. In a case of RDKit [29], we reported the performance from the generated conformers from RDKit that we utilized to compute the approximate fragment coordinates $\hat{\mathbf{x}}^{\mathrm{f}} \sim p_{\mathrm{RDKit}}(\mathbf{x}^{\mathrm{f}})$. For GeoDiff [55], we downloaded their implementation code from `https://github.com/MinkaiXu/GeoDiff/tree/main` and trained GeoDiff model for our experiments. We reported the performance of GeoDiff after sampling conformers using Langevin dynamics [48], as they did in their implementation.

E.5 Pseudo-code

In this subsection, we provide the Pytorch-style [35] pseudo-codes. The RDKit conformer generator to obtain the approximate fragment structure, linear interpolation blurring schedule, training process, and sampling process were given in Pseudo-codes 1, 2, 3, and 4, respectively.

```
import torch
import copy
from rdkit.Chem import AllChem

def get_multiple_rdkit_coords(molecule, num_conf):
    mol = copy.deepcopy(molecule)
    mol.RemoveAllConformers()
    ps = AllChem.ETDG()
    ps.maxIterations = 5000
    ps.randomSeed = 2023
    ps.useBasicKnowledge = False
    ps.useExpTorsionAnglePrefs = False
    ps.useRandomCoords = False
    ids = AllChem.EmbedMultipleConfs(mol, num_conf, ps)
    if -1 in ids or mol.GetNumConformers() != num_conf:
        print("Use DG random coords.")
        ps.useRandomCoords = True
        ids = AllChem.EmbedMultipleConfs(mol, num_conf, ps)
    confs = []
    for cid in range(num_conf):
        confs.append(torch.tensor(mol.GetConformer(cid).GetPositions()
    )

    return confs
```

Pseudo-code 1: Initial atom coordinate generation from RDKit.

```python
import torch

def blurring(t, x_a_gt, x_f_rdkit, mapping_matrix):
    # prior distribution
    x_f_rdkit_extend = mapping_matrix @ x_f_rdkit

    # move positions to zero center-of-mass subspace
    x_a_gt = remove_mean(x_a_gt)
    x_f_rdkit_extend = remove_mean(x_f_rdkit_extend)

    # linear interpolation
    blurred_pos = torch.lerp(x_a_gt, x_f_ref_ext_split, t)

    return blurred_pos
```

Pseudo-code 2: Blurring schedule in Eq. 2.

```python
import torch

def loss(x_a_gt, x_f_rdkit, mapping_matrix, sigma, T):
    # sample time
    t = torch.randint(1, T, (1,)) / T

    # blurred atom position from blurring schedule
    blurred_pos = blurring(t, x_a_gt, x_f_rdkit, mapping_matrix)

    # add noise
    noise = torch.randn_like(blurred_pos)
    noise = remove_mean(noise)
    blurred_pos = blurred_pos + noise * sigma

    # estimate ground truth state from blurred atom position
    x_a_gt_estimated = deblur_network(blurred_pos, mapping_matrix, t)

    # translate to the zero center-of-mass subspace
    x_a_gt = remove_mean(x_a_gt)
    x_a_gt_est = remove_mean(x_a_gt_estimated)

    # optimal rotation matrix from Kabsch algorithm
    rot_matrix = Kabsch_alignment(x_a_gt_est, x_a_gt)

    # mean squared error
    loss = mean((x_a_gt - rot_matrix @ x_a_gt_est) ** 2)

    return loss
```

Pseudo-code 3: Training process.

```python
import torch
import copy

def sample(x_f_rdkit, mapping_matrix, delta, T):
    # initial atom position located at fragment position
    x_a_init = mapping_matrix @ x_f_rdkit
    x_a_init = remove_mean(x_a_init)
    x_a = copy.deepcopy(x_a_init)

    for i in range(T-1, 0, -1):
        t = i/T

        # add noise
        noise = torch.randn_like(x_a)
        noise = remove_mean(noise)
        x_a = x_a + noise * delta

```

```
        # estimate ground truth state from blurred atom position
        x_a_gt_est = deblur_network(x_a, mapping_matrix, t)

        # translate to the zero center-of-mass subspace
        x_a_gt_est = remove_mean(x_a_gt_est)

        # optimal rotation matrix from Kabsch algorithm
        rot_matrix = Kabsch_alignment(x_a_gt_est, x_a_init)

        # next step from estimated ground truth and initial positions
        x_a_gt_est = rot_matrix @ x_a_gt_est
        x_a = blurring((i-1)/T, x_a_gt_est, x_a_init, mapping_matrx)

    return x_a
```

Pseudo-code 4: Sampling process.

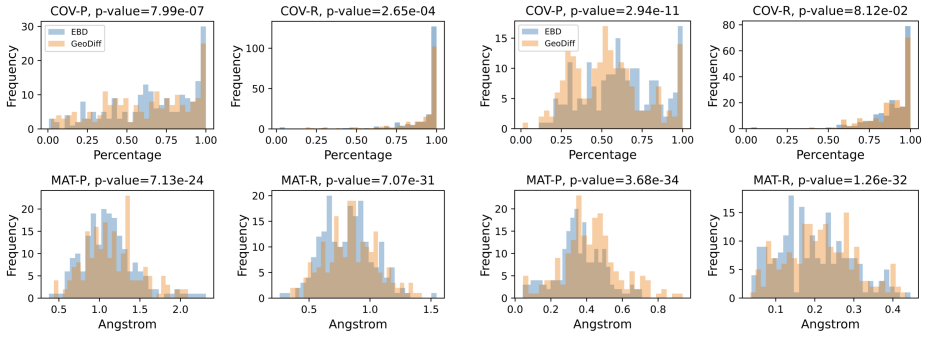
F Further results on geometric evaluation

GEOM-QM9. We compared our EBD to the baseline RDKit and machine learning models on small molecules GEOM-QM9, and the results are reported in Table 6. Compared to the most of machine learning models, EBD achieved superior performances especially on the precision score. We observed that RDKit, the distance geometry-based conformer generator, outperformed in coverage metrics for small molecules. However, as the size of molecules increases and the tasks become more challenging, RDKit suffers a significant performance drop, as shown in Table 2.

Table 6: Geometric evaluation on GEOM-QM9 benchmark ($\delta = 0.5$Å).

Models	COV-R (%) ↑		MAT-R(Å) ↓		COV-P (%) ↑		MAT-P (Å) ↓	
	Mean	Med	Mean	Med	Mean	Med	Mean	Med
RDKit	88.34	**95.08**	0.3544	0.2974	**83.42**	**88.17**	0.3747	0.3692
CVGAE	0.09	0.00	1.6713	1.6088	-	-	-	-
GraphDG	73.33	84.21	0.4245	0.3973	43.90	35.33	0.5809	0.5823
CGCF	78.05	82.48	0.4219	0.3900	36.49	33.57	0.6615	0.6427
ConfVAE	77.84	88.20	0.4154	0.3739	38.02	34.67	0.6215	0.6091
GeoMol	71.26	72.00	0.3731	0.3731	-	-	-	-
ConfGF	88.49	94.31	0.2673	0.2685	46.43	43.41	0.5224	0.5124
GeoDiff ($T = 5000$)	88.02	92.33	**0.2199**	0.2116	53.72	52.36	0.4362	0.4259
EBD ($T = 50$)	**89.37**	93.21	0.2374	**0.1903**	61.31	60.46	**0.3622**	**0.3517**

Statistical significance. We report the statistical significance of our model's improvements in geometric evaluation (COV-P, COV-R, MAT-P, and MAT-R scores). We measured p-value from one-sided Wilcoxon signed-rank test (a non-parametric version of paired t-test) over those scores of EBD and GeoDiff [55] on Drugs and QM9, and the results are reported in Fig. 5. Except for the COV-R score on QM9, our EBD achieved statistically significant improvement in generating more diverse and more accurate conformers for every score on either dataset, as evidenced by the p-value.



(a) Drugs (b) QM9

Figure 5: p-value of COV-P, COV-R, MAT-P, and MAT-R on Drugs and QM9.

G Visualizations

We provide additional samples and sampling processes of EBD for the test set of Drugs in Figs. 6, 8 and the test set of QM9 in Figs. 7, 9.

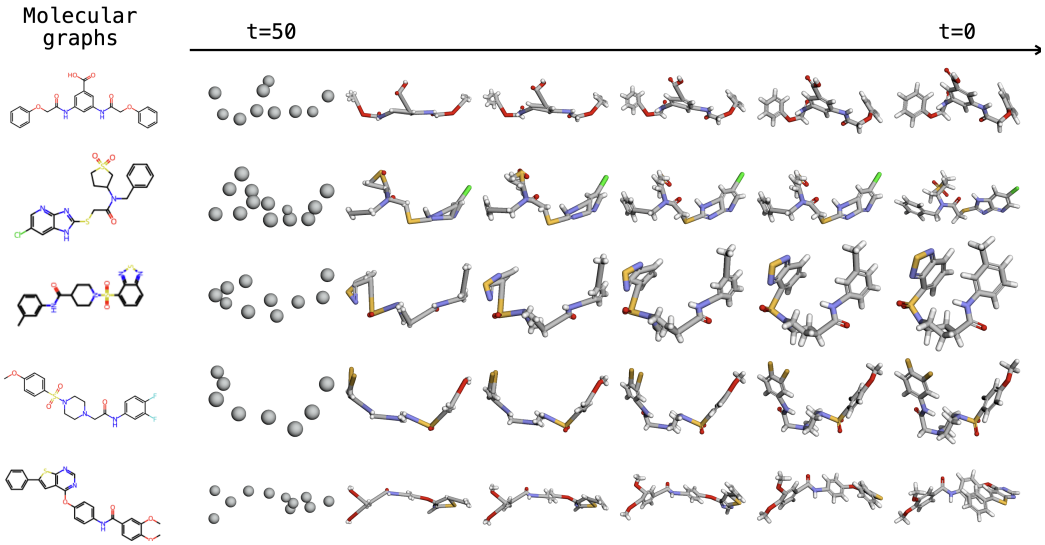


Figure 6: Sampling processes of EBD on Drugs.

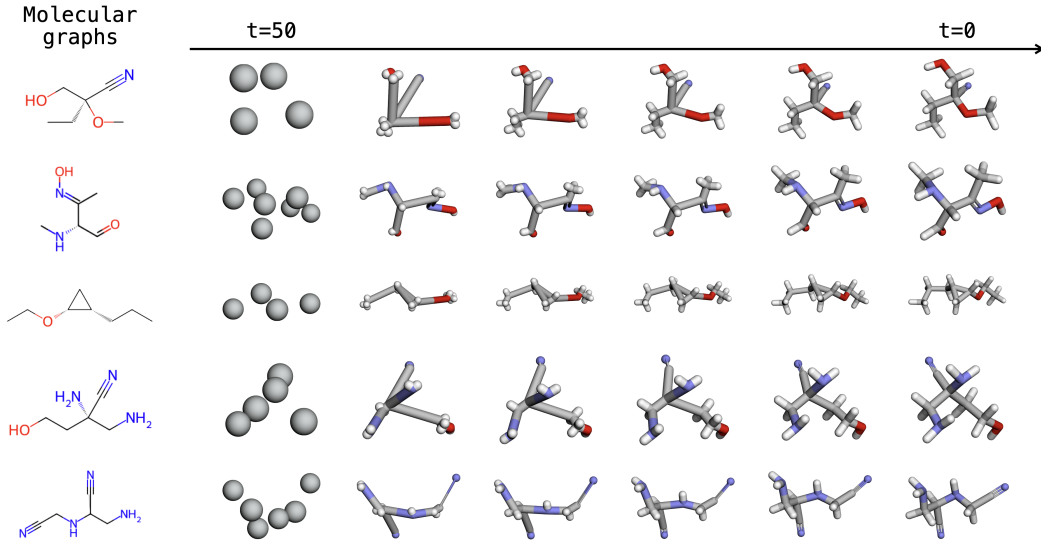


Figure 7: Sampling processes of EBD on QM9.

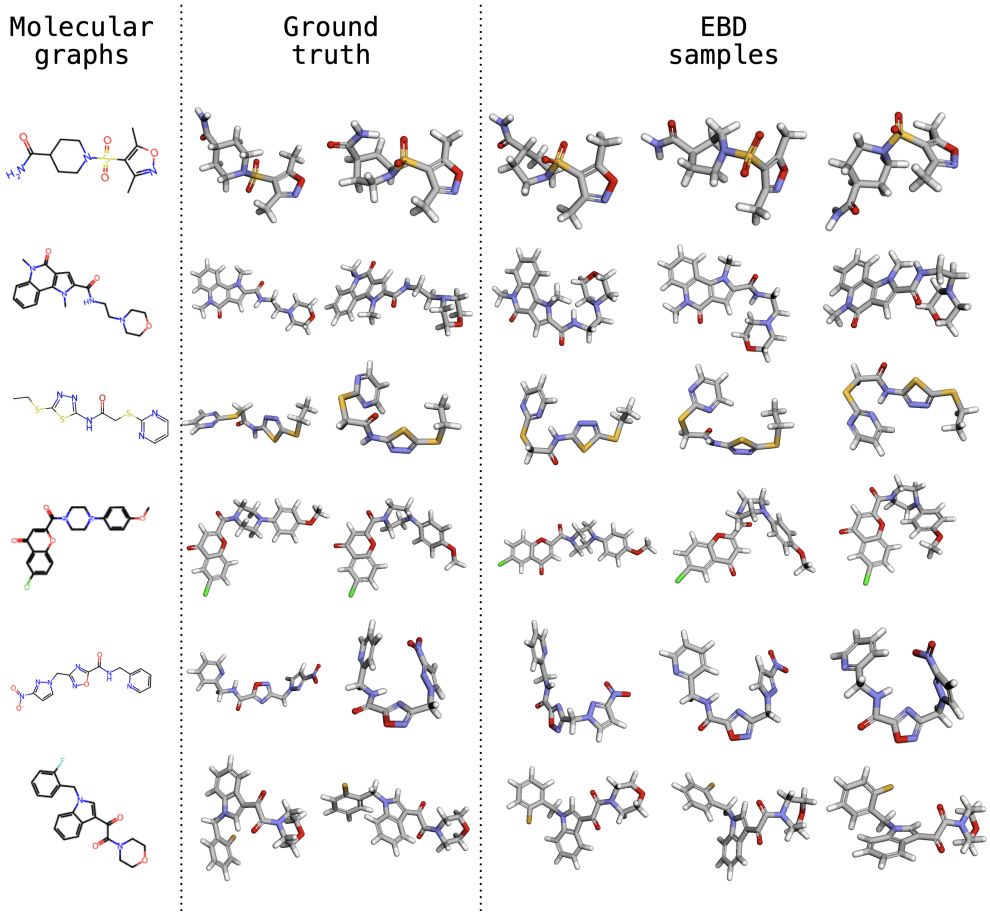


Figure 8: Visualization of molecular graphs, ground truth conformers, and samples of EBD on Drugs.

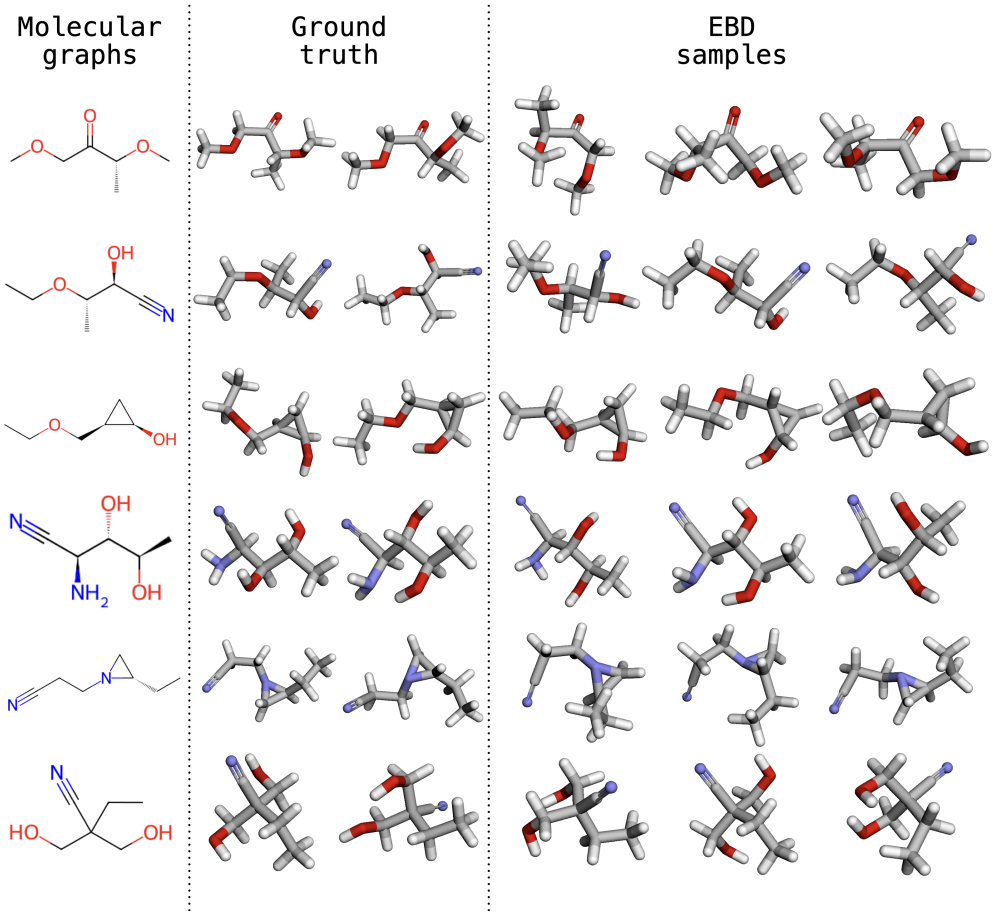


Figure 9: Visualization of molecular graphs, ground truth conformers, and samples of EBD on QM9.