ABSTRACT

Predicting molecular conformations from molecular graphs is a fundamental problem in cheminformatics and drug discovery. Recently, significant progress has been achieved with machine learning approaches, especially with deep generative models. Inspired by the diffusion process in classical non-equilibrium thermodynamics where heated particles will diffuse from original states to a noise distribution, in this paper, we propose a novel generative model named GeoDiff for molecular conformation prediction. GeoDiff treats each atom as a particle and learns to directly reverse the diffusion process (i.e., transforming from a noise distribution to stable conformations) as a Markov chain. Modeling such a generation process is however very challenging as the likelihood of conformations should be roto-translational invariant. We theoretically show that Markov chains evolving with equivariant Markov kernels can induce an invariant distribution by design, and further propose building blocks for the Markov kernels to preserve the desirable equivariance property. The whole framework can be efficiently trained in an end-to-end fashion by optimizing a weighted variational lower bound to the (conditional) likelihood. Experiments on multiple benchmarks show that GeoDiff is superior or comparable to existing state-of-the-art approaches, especially on large molecules.  

1 INTRODUCTION

Graph representation learning has achieved huge success for molecule modeling in various tasks ranging from property prediction (Gilmer et al., 2017; Duvenaud et al., 2015) to molecule generation (Jin et al., 2018; Shi et al., 2020), where typically a molecule is represented as an atom-bond graph. Despite its effectiveness in various applications, a more intrinsic and informative representation for molecules is the 3D geometry, also known as conformation, where atoms are represented as their Cartesian coordinates. The 3D structures determine the biological and physical properties of molecules and hence play a key role in many applications such as computational drug and material design (Thomas et al., 2018; Gebauer et al., 2021; Jing et al., 2021; Batzner et al., 2021). Unfortunately, how to predict stable molecular conformation remains a challenging problem. Traditional methods based on molecular dynamics (MD) or Markov chain Monte Carlo (MCMC) are very computationally expensive, especially for large molecules (Hawkins, 2017).

Recently, significant progress has been made with machine learning approaches, especially with deep generative models. For example, Simm & Hernandez-Lobato (2020); Xu et al. (2021b) studied predicting atomic distances with variational autoencoders (VAEs) (Kingma & Welling, 2013) and flow-based models (Dinh et al., 2017) respectively. Shi et al. (2021) proposed to use denoising score matching (Song & Ermon, 2019; 2020) to estimate the gradient fields over atomic distances, through which the gradient fields over atomic coordinates can be calculated. Ganea et al. (2021) studied generating conformations by predicting both bond lengths and angles. As molecular conformations are roto-translational invariant, these approaches circumvent directly modeling atomic coordinates by leveraging intermediate geometric variables such as atomic distances, bond and torsion angles, which
are roto-translational invariant. As a result, they are able to achieve very compelling performance. However, as all these approaches seek to indirectly model the intermediate geometric variables, they have inherent limitations in either training or inference process (see Sec. 2 for a detailed description). Therefore, an ideal solution would still be directly modeling the atomic coordinates and at the same time taking the roto-translational invariance property into account.

In this paper, we propose such a solution called GEODIFF, a principled probabilistic framework based on denoising diffusion models (Sohl-Dickstein et al., 2015). Our approach is inspired by the diffusion process in nonequilibrium thermodynamics (De Groot & Mazur, 2013). We view atoms as particles in a thermodynamic system, which gradually diffuse from the original states to a noisy distribution in contact with a heat bath. At each time step, stochastic noises are added to the atomic positions. Our high-level idea is learning to reverse the diffusion process, which recovers the target geometric distribution from the noisy distribution. In particular, inspired by recent progress of denoising diffusion models on image generation (Ho et al., 2020; Song et al., 2020), we view the noisy geometries at different timesteps as latent variables, and formulate both the forward diffusion and reverse denoising process as Markov chains. Our goal is to learn the transition kernels such that the reverse process can recover realistic conformations from the chaotic positions sampled from a noise distribution. However, extending existing methods to geometric generation is highly non-trivial: a direct application of diffusion models on the conformation generation task lead to poor generation quality. As mentioned above, molecular conformations are roto-translational invariant, i.e., the estimated (conditional) likelihood should be unaffected by translational and rotational transformations (Köhler et al., 2020). To this end, we first theoretically show that a Markov process starting from an roto-translational invariant prior distribution and evolving with roto-translational equivariant Markov kernels can induce an roto-translational invariant density function. We further provide practical parameterization to define a roto-translational invariant prior distribution and a Markov kernel imposing the equivariance constraints. In addition, we derive a weighted variational lower bound of the conditional likelihood of molecular conformations, which also enjoys the roto-translational invariance and can be efficiently optimized.

A unique strength of GEODIFF is that it directly acts on the atomic coordinates and entirely bypasses the usage of intermediate elements for both training and inference. This general formulation enjoys several crucial advantages. First, the model can be naturally trained end-to-end without involving any sophisticated techniques like bilevel programming (Xu et al., 2021b), which benefits from small optimization variances. Besides, instead of solving geometries from bond lengths or angles, the one-stage sampling fashion avoids accumulating any intermediate error, and therefore leads to more accurate predicted structures. Moreover, GEODIFF enjoys a high model capacity to approximate the complex distribution of conformations. Thus, the model can better estimate the highly multi-modal distribution and generate structures with high quality and diversity.

We conduct comprehensive experiments on multiple benchmarks, including conformation generation and property prediction tasks. Numerical results show that GEODIFF consistently outperforms existing state-of-the-art machine learning approaches, and by a large margin on the more challenging large molecules. The significantly superior performance demonstrate the high capacity to model the complex distribution of molecular conformations and generate both diverse and accurate molecules.

## 2 RELATED WORK

Recently, various deep generative models have been proposed for conformation generation. Among them, CVGAE (Mansimov et al., 2019) first proposed a VAE model to directly generate 3D atomic coordinates, which fails to preserve the roto-translation equivariance property of conformations and suffers from poor performance. To address this problem, the majority of subsequent models are based on intermediate geometric elements such as atomic distances and torsion angles. A favorable property of these elements is the roto-translational invariance, (e.g. atomic distances does not change when rotating the molecule), which has been shown to be an important inductive bias for molecular geometry modeling (Köhler et al., 2020). However, such a decomposition suffers from several drawbacks for either training or sampling. For example, GRAPHDG (Simm & Hernandez-Lobato, 2020) and CGCF (Xu et al., 2021a) proposed to predict the interatomic distance matrix by VAE and Flow respectively, and then solve the geometry through the Distance Geometry (DG) technique (Liberti et al., 2014), which searches reasonable coordinates that matches with the predicted
distances. ConfVAE further improves this pipeline by designing an end-to-end framework via bilevel optimization (Xu et al., 2021b). However, all these approaches suffer from the accumulated error problem, meaning that the noise in the predicted distances will misguide the coordinate searching process and lead to inaccurate or even erroneous structures. To overcome this problem, ConfGF (Shi et al., 2021; Luo et al., 2021) proposed to learn the gradient of the log-likelihood w.r.t coordinates. However, in practice the model is still aided by intermediate geometric elements, in that it first estimates the gradient w.r.t interatomic distances via denoising score matching (DSM) (Song & Ermon, 2019; 2020), and then derives the gradient of coordinates using the chain rule. The problem is, by learning the distance gradient via DSM, the model is fed with perturbed distance matrices, which may violate the triangular inequality or even contain negative values. As a consequence, the model is actually learned over invalid distance matrices but tested with valid ones calculated from coordinates, making it suffer from serious out-of-distribution (Hendrycks & Gimpel, 2016) problem. Most recently, another concurrent work (Ganea et al., 2021) proposed a highly systematic (rule-based) pipeline named GeoMol, which learns to predict a minimal set of geometric quantities (i.e. length and angles) and then reconstruct the local and global structures of the conformation in a sophisticated procedure. Besides, there has also been efforts to use reinforcement learning for conformation search Gogineni et al. (2020). Nevertheless, this method relies on rigid rotor approximation and can only model the torsion angles, and thus fundamentally differs from other approaches.

3 Preliminaries

3.1 Notations and Problem Definition

Notations. In this paper each molecule with $n$ atoms is represented as an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_i\}_{i=1}^n$ is the set of vertices representing atoms and $\mathcal{E} = \{e_{ij} \mid (i, j) \subseteq |\mathcal{V}| \times |\mathcal{V}|\}$ is the set of edges representing inter-atomic bonds. Each node $v_i \in \mathcal{V}$ describes the atomic attributes, e.g., the element type. Each edge $e_{ij} \in \mathcal{E}$ describes the corresponding connection between $v_i$ and $v_j$, and is labeled with its chemical type. In addition, we also assign the unconnected edges with a virtual type. For the geometry, each atom in $\mathcal{V}$ is embedded by a coordinate vector $c \in \mathbb{R}^3$ into the 3-dimensional space, and the full set of positions (i.e., the conformation) can be represented as a matrix $C = [c_1, c_2, \cdots, c_n] \in \mathbb{R}^{n \times 3}$.

Problem Definition. The task of molecular conformation generation is a conditional generative problem, where we are interested in generating stable conformations for a provided graph $G$. Given multiple graphs $G$, and for each $G$ given its conformations $C$ as i.i.d samples from an underlying Boltzmann distribution (Noé et al., 2019), our goal is learning a generative model $p_\theta(C|G)$, which is easy to draw samples from, to approximate the Boltzmann function.

3.2 Equivariance

Equivariance is ubiquitous in machine learning for atomic systems, e.g., the vectors of atomic dipoles or forces should rotate accordingly w.r.t. the conformation coordinates (Thomas et al., 2018; Weiler et al., 2018; Fuchs et al., 2020; Miller et al., 2020; Simm et al., 2021; Batzner et al., 2021). It has shown effectiveness to integrate such inductive bias into model parameterization for modeling 3D geometry, which is critical for the generalization capacity (Köhler et al., 2020; Satorras et al., 2021a). Formally, a function $F: X \rightarrow Y$ is equivariant w.r.t a group $G$ if:

$$F \circ T_g (x) = S_g \circ F(x),$$

(1)

where $T_g$ and $S_g$ are transformations for an element $g \in G$, acting on the vector spaces $X$ and $Y$, respectively. In this work, we consider the SE(3) group, i.e., the group of rotation, translation in 3D space. This requires the estimated likelihood unaffected with translational and rotational transformations, and we will elaborate on how our method satisfy this property in Sec. 4.

4 GeoDiff Method

In this section, we elaborate on the proposed equivariant diffusion framework. We first present a high level description of our 3D diffusion formulation in Sec. 4.1, based on recent progress of denoising diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020). Then we emphasize several
4.1 FORMULATION

Let $C^0$ denotes the ground truth conformations and let $C^t$ for $t = 1, \ldots, T$ be a sequence of latent variables with the same dimension, where $t$ is the index for diffusion steps. Then a diffusion probabilistic model (Sohl-Dickstein et al., 2015) can be described as a latent variable model with two processes: the forward diffusion process, and the reverse generative process. Intuitively, the diffusion process progressively injects small noises to the data $C^0$, while the generative process learns to revert the diffusion process by gradually eliminating the noise to recover the ground truth. We provide a high-level schematic of the processes in Fig. 1.

**Diffusion process.** Following the physical insight, we model the particles $C$ as an evolving thermodynamic system. With time going by, the equilibrium conformation $C^0$ will gradually diffuse to the next chaotic states $C^t$, and finally converge into a white noise distribution after $T$ iterations. Different from typical latent variable models, in diffusion model this forward process is defined as a fixed (rather than trainable) posterior distribution $q(C^1:T|C^0)$. Specifically, we define it as a Markov chain according to a fixed variance schedule $\beta_1, \ldots, \beta_T$:

$$q(C^1:T|C^0) = \prod_{t=1}^{T} q(C^t|C^{t-1}), \quad q(C^t|C^{t-1}) = \mathcal{N}(C^t; \sqrt{1 - \beta_t}C^{t-1}, \beta_t I).$$  

Note that, in this work we do not impose specific (invariance) requirement upon the diffusion process, as long as it can efficiently draw noisy samples for training the generative process $p_0(C^0)$.

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, a special property of the forward process is that $q(C^t|C^0)$ of arbitrary timestep $t$ can be calculated in closed form $q(C^t|C^0) = \mathcal{N}(C^t; \sqrt{\bar{\alpha}_t}C^0, (1 - \bar{\alpha}_t)I)^2$. This indicates with sufficiently large $T$, the whole forward process will convert $C^0$ to whitened isotropic Gaussian, and thus it is natural to set $p(C^T)$ as a standard Gaussian distribution.

**Reverse Process.** Our goal is learning to recover conformations $C^0$ from the white noise $C^T$, given specified molecular graphs $G$. We consider this generative procedure as a reverse dynamics of the above diffusion process, starting from the noisy particles $C^T \sim p(C^T)$. We formulate this reverse dynamics as a conditional Markov chain with learnable transitions:

$$p_0(C^{0:T-1}|G, C^T) = \prod_{t=1}^{T} p_0(C^{t-1}|G, C^t), \quad p_0(C^{t-1}|G, C^t) = \mathcal{N}(C^{t-1}; \mu_0(G, C^t, t), \sigma_t^2 I).$$  

Herein $\mu_0$ are parameterized neural networks to estimate the means, and $\sigma_t$ can be any user-defined variance. The initial distribution $p(C^T)$ is set as a standard Gaussian. Given a graph $G$, its 3D structure is generated by first drawing chaotic particles $C^T$ from $p(C^T)$, and then iteratively refined through the reverse Markov kernels $p_0(C^{t-1}|G, C^t)$.

\[2\] Detailed derivations are provided in the Appendix A.
Having formulated the reverse dynamics, the marginal likelihood can be calculated by \( p_0(C^0|G) = \int p(C^T)p_0(C^{t:T-1}|G,C^t)\,dC^{1:T} \). Herein a non-trivial problem is that the likelihood should be invariant w.r.t. translation and rotation, which has proved to be a critical inductive bias for 3D object generation (Köhler et al., 2020; Satorras et al., 2021a). In the following subsections, we will elaborate on how we parameterize the Markov kernels \( p_0(C^{t-1}|G,C^t) \) to achieve this desired property, and also how to maximize this likelihood by taking the invariance into account.

4.2 Equivariant Reverse Generative Process

Instead of directly leveraging existing methods, we consider building the density \( p_0(C^0) \) that is invariant to rotation and translation transformations. Intuitively, this requires the likelihood to be unaffected by translations and rotations. Formally, let \( T_g \) be some roto-translational transformations of a group element \( g \in \text{SE}(3) \), then we have the following statement:

**Proposition 1.** Let \( p(x_T) \) be an \( \text{SE}(3) \)-invariant density function, i.e., \( p(x_T) = p(T_g(x_T)) \). If Markov transitions \( p(x_{t-1}|x_t) \) are \( \text{SE}(3) \)-equivariant, i.e., \( p(x_{t-1}|x_t) = p(T_g(x_{t-1})|T_g(x_t)) \), then we have that the density \( p_0(x_0) = \int p(x_T)p_0(x_{0:T-1}|x_T)\,dx_{1:T} \) is also \( \text{SE}(3) \)-invariant.

This proposition indicates that the dynamics starting from an invariant standard density along an equivariant Gaussian Markov kernel can result in an invariant density. Now we provide a practical implementation of GEO_DIFF based on the recent denoising diffusion framework (Ho et al., 2020).

**Invariant Initial Density** \( p(C^T) \). We first introduce the invariant distribution \( p(C^T) \), which will also be employed in the equivariant Markov chain. We borrow the idea from Köhler et al. (2020) to consider systems with zero center of mass (CoM), termed CoM-free systems. We define \( p(C^T) \) as a “CoM-free standard density” \( \rho(C) \), built upon an isotropic normal density \( \rho(C) \): for evaluating the likelihood \( \tilde{\rho}(C) \) we can firstly translate \( C \) to zero CoM and then calculate \( \rho(C) \), and for sampling from \( \tilde{\rho}(C) \) we can firstly sample from \( \rho(C) \) and then move the CoM to zero.

We provide a formal theoretical analysis of \( \tilde{\rho}(C) \) in Appendix A. Intuitively, the isotropic Gaussian is manifestly invariant to rotations around the zero CoM. And by considering CoM-free system, moving the particles to zero CoM can always ensure the translational invariance. Consequently, \( \tilde{\rho}(C) \) is constructed as a roto-transitional invariant density.

**Equivariant Markov Kernels** \( p(C^{t-1}|G,C^t) \). Similar to the prior density, we also consider equipping all intermediate structures \( C^t \) as CoM-free systems. Specifically, given mean \( \mu_0(G,C^t,t) \) and variance \( \sigma_t \), the likelihood of \( C^{t-1} \) will be calculated by \( \tilde{\rho}(C^{t-1}|G,C^t) = \frac{1}{\sigma_t} \rho(C^{t-1} - \mu_0(G,C^t,t)) \). The CoM-free Gaussian ensures the translation invariance in the Markov kernels. Consequently, to achieve the equivariant property defined in Proposition 1, we focus on the rotation equivariance.

Then in general, the key requirement is to ensure the means \( \mu_0(G,C^t,t) \) to be roto-translation equivariant w.r.t. \( C^t \). Following Ho et al. (2020), we consider the following parameterization of \( \mu_0 \):

\[
\mu_0(C^t, t) = \frac{1}{\sqrt{\alpha_t}} \left( C^t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(G,C^t, t) \right),
\]

where \( \epsilon_\theta \) are neural networks with trainable parameters \( \theta \). Intuitively, the model \( \epsilon_\theta \) learns to predict the noise necessary to decorrupt the conformations. This is analogous to the physical force fields (Schütt et al., 2017; Zhang et al., 2018; Hu et al., 2021; Shuaibi et al., 2021), which also gradually push particles towards convergence around the equilibrium states.

Now the problem is transformed to constructing \( \epsilon_\theta \) to be roto-translational equivariant. We draw inspirations from recent equivariant networks (Thomas et al., 2018; Satorras et al., 2021b) to design an equivariant convolutional layer, named graph field network (GFN). In the \( l \)-th layer, GFN takes node embeddings \( \mathbf{h}_l \in \mathbb{R}^{n \times b} \) (\( b \) denotes the feature dimension) and corresponding coordinate embeddings \( \mathbf{x}_l \in \mathbb{R}^{n \times 3} \) as inputs, and outputs \( \mathbf{h}_{l+1} \) and \( \mathbf{x}_{l+1} \) as follows:

\[
\mathbf{m}_{ij} = \Phi_m(\mathbf{h}_l^i, \mathbf{h}_l^j, \| \mathbf{x}_l^i - \mathbf{x}_l^j \|^2, \epsilon_{ij}; \theta_m)
\]

\[
\mathbf{h}_{l+1}^i = \Phi_h(\mathbf{h}_l^i, \sum_{j \in N(i)} \mathbf{m}_{ij}; \theta_h)
\]

\[
\mathbf{x}_{l+1}^i = \sum_{j \in N(i)} \frac{1}{d_{ij}} (\mathbf{c}_i - \mathbf{c}_j) \Phi_x(\mathbf{m}_{ij}; \theta_x)
\]
where $\Phi$ are feed-forward networks and $d_{ij}$ denotes interatomic distances. $\mathcal{N}(i)$ denotes the neighborhood of $i^{th}$ node, including both connected atoms and other ones within a radius threshold $\tau$, which enables the model to explicitly capture long-range interactions and support molecular graphs with disconnected components. Initial embeddings $h^0$ are combinations of atom and timestep embeddings, and $x^0$ are atomic coordinates. The main difference between proposed GFN and other GNNs lies in equation 7, where $x$ is updated as a combination of radial directions weighted by $\Phi_x : \mathbb{R}^b \rightarrow \mathbb{R}$.

Such vector field $x^c$ enjoys the roto-translation equivariance property. Formally, we have:

**Proposition 2.** Parameterizing $\epsilon_0(G, C, t)$ as a composition of $L$ GFN layers, and take the $x^c$ after $L$ updates as the output. Then the noise vector field $\epsilon_0$ is SE(3) equivariant w.r.t the 3D system $C$.

Intuitively, given $h^t$ already invariant and $x^t$ equivariant, the message embedding $m$ will also be invariant since it only depends on invariant features. Since $x$ is updated with the relative differences $c_i - c_j$ weighted by invariant features, it will be translation-invariant and rotation-equivariant. Then inductively, composing $\epsilon_0$ with $L$ GFN layers enables equivariance with $C^t$. We provide the formal proof of equivariance properties in Appendix A.

### 4.3 Improved Training Objective

Having formulated the generative process and the model parameterization, now we consider the practical training objective for the reverse dynamics. Since directly optimizing the exact log-likelihood is intractable, we instead maximize the usual variational lower bound (ELBO):

$$
\mathbb{E} \left[ \log p_0(C^0|G) \right] = \mathbb{E} \left[ \log \mathbb{E}_{q(C^{t-1}|C^0)} p_0(C^t|C^0) q(G|C^0) \right] \\
\geq -\mathbb{E}_{q} \left[ \sum_{t=1}^{T} D_{KL}(q(C^{t-1}|C^0, C^0) \| p_0(C^{t-1}|C^t, G)) \right] := -\mathcal{L}_{\text{ELBO}}
$$

where $q(C^{t-1}|C^t, C^0)$ is analytically tractable as $\mathcal{N}(\sqrt{\frac{\alpha_t-1}{\alpha_t}} C^0 + \sqrt{\frac{\alpha_t(1-\alpha_t-1)}{1-\alpha_t}} C^t, \frac{1-\alpha_t-1}{1-\alpha_t} \beta_t)^3$. Most recently, Ho et al. (2020) showed that under the parameterization in equation 4, the ELBO of the diffusion model can be further simplified by calculating the KL divergences between Gaussians as weighted $L_2$ distances between the means $\epsilon_0$ and $c^t$. Formally, we have:

**Proposition 3.** (Ho et al., 2020) Under the parameterization in equation 4, we have:

$$
\mathcal{L}_{\text{ELBO}} = \sum_{t=1}^{T} \gamma_t \mathbb{E}_{G, \epsilon \sim \mathcal{N}(0, 1)} \left[ \| \epsilon - \epsilon_0(G, C^t, t) \|_2^2 \right]
$$

where $C^t = \sqrt{\alpha_t} C^0 + \sqrt{1-\alpha_t} \epsilon$. The weights $\gamma_t = \frac{\beta_t}{2 \alpha_t (1-\alpha_t-1)}$ for $t > 1$, and $\gamma_1 = \frac{1}{2 \alpha_1}$.

The intuition of this objective is to independently sample chaotic conformations of different timesteps from $q(C^{t-1}|C^t, C^0)$, and use $\epsilon_0$ to model the noise vector $\epsilon$. To yield a better empirical performance, Ho et al. (2020) suggests to set all weights $\gamma_t$ as 1, which is in line with the objectives of recent noise conditional score networks (Song & Ermon, 2019, 2020).

As $\epsilon_0$ is designed to be equivariant, it is natural to require its supervision signal $\epsilon$ to be equivariant with $C^t$. Note that once this is achieved, the ELBO will also become invariant. However, the $\epsilon$ in the forward diffusion process is not imposed with such equivariance, violating the above properties. Here we propose two approaches to obtain the modified noise vector $\hat{\epsilon}$, which, after replacing $\epsilon$ in the $L_2$ distance calculation in equation 9, achieves the desired equivariance:

**Alignment approach.** Considering the fact that $\epsilon$ can be calculated by $\frac{C^{t-1} - \sqrt{\alpha_t} C^t}{\sqrt{1-\alpha_t}}$, we can first rotate and translate $C^0$ to $\hat{C}^0$ by aligning w.r.t $C^t$, and then compute $\hat{\epsilon}$ as $\frac{\hat{C}^t - \sqrt{\alpha_t} C^t}{\sqrt{1-\alpha_t}}$. Since the aligned conformation $\hat{C}^0$ is equivariant with $C^t$, the processed $\hat{\epsilon}$ will also enjoy the equivariance. Specifically, the alignment is implemented by first translating $C^0$ to the same CoM of $C^t$ and then solve the optimal rotation matrix by Kabsch alignment algorithm (Kabsch, 1976).

---

1The detailed derivations and full proofs are provided in Appendix A.
Chain-rule approach. Another meaningful observation is that by reparameterizing the Gaussian
distribution \( q(C^t | C^0) \) as \( C^t = \sqrt{\alpha_t} C^0 + \sqrt{1 - \alpha_t} \epsilon, \epsilon \) can be viewed as a weighted score function
\( \sqrt{1 - \alpha_t} \nabla_C q(C^t | C^0) \). Shi et al. (2021) recently shows that generally this score function \( \nabla_C q(C^t | \cdot) \)
can be designed to be equivariant by decomposing it into \( \partial_C \epsilon \) \( \nabla_C q(C^t | \cdot) \) with the chain rule, where
\( \epsilon \) can be any invariant features of the structures \( C^t \) such as the inter-atomic distances. We refer
readers to Shi et al. (2021) for more details. The insight is that as gradient of invariant variables \( w.r.t \)
equivariant variables, the partial derivative \( \partial_C \epsilon \) will always be equivalent with \( C^t \). In this work, under the common assumption that \( \epsilon \) also follows a Gaussian distribution (Kingma & Welling, 2013),
our practical implementation is to first approximately calculate \( \nabla_C q(C^t | C^0) \) as \( \frac{\partial_t \epsilon - \sqrt{\alpha_t} \epsilon 0}{1 - \alpha_t} \), and then compute the modified noise vector \( \epsilon \) as \( \sqrt{1 - \alpha_t} \partial_C \epsilon \left( \frac{\partial_t \epsilon - \sqrt{\alpha_t} \epsilon 0}{1 - \alpha_t} \right) = \frac{\partial_t \epsilon \epsilon 0}{\sqrt{1 - \alpha_t}} \).

4.4 Sampling

With a learned reverse dynamics \( \epsilon_\theta(G, C^t, t) \), the transition means \( \mu_\theta(G, C^t, t) \) can be calculated by equation 4. Thus, given a graph
\( G \), its geometry \( C^0 \) is generated by first sampling chaotic particles \( C^T \sim p(C^T) \), and then progressively sample \( C^{t-1} \sim p_\theta(C^{t-1} | G, C^t) \) for \( t = T, T - 1, \ldots, 1 \). This process is Markovian, which gradually shifts the previous noisy positions towards equilibrium states. We provide the pseudo code of the whole sampling process in Algorithm 1.

5 EXPERIMENT

In this section, we empirically evaluate GEODIFF on the task of equilibrium conformation generation
for both small and drug-like molecules. Following existing work (Shi et al., 2021; Ganea et al.,
2021), we test the proposed method as well as the competitive baselines on two standard benchmarks:
Conformation Generation (Sec. 5.2) and Property Prediction (Sec. 5.3). We first present the
general experiment setups, and then describe task-specific evaluation protocols and discuss the results
in each section. The implementation details are provided in Appendix C.

5.1 EXPERIMENT SETUP

Datasets. Following prior works (Xu et al., 2021a; b), we also use the recent GEOM-QM9 (Ramakrishnan et al., 2014) and GEOM-Drugs (Axeled & Gomez-Bombarelli, 2020) datasets. The former
one contains small molecules while the latter one are medium-sized organic compounds. We borrow
the data split produced by Shi et al. (2021). For both datasets, the training split consists of 40,000 molecules with 5 conformations for each, resulting in 200,000 conformations in total. The valid split share the same size as training split. The test split contains 200 distinct molecules, with 22, 408 conformations for QM9 and 14, 324 ones for Drugs.

Baseline. We compare GEODIFF with 6 recent or established state-of-the-art baselines. For the ML
approaches, we test the following models with highest reported performance: CVGAE (Mansimov et al., 2019), GRAPHGD (Simm & Hernandez-Lobato, 2020), CGCF (Xu et al., 2021a), CONF-
VAE (Xu et al., 2021b) and CONF-GF (Shi et al., 2021). We also test the classic RDKit (Riniker & Landrum, 2015) method, which is arguably the most popular open-source software for conformation generation. We refer readers to Sec. 2 for a detailed discussion of these models.

5.2 CONFORMATION GENERATION

Evaluation metrics. The task aims to measure both quality and diversity of generated conformations
by different models. We follow Ganea et al. (2021) to evaluate 4 metrics built upon root-mean-square

![Algorithm 1 Sampling Algorithm of GEODIFF.](image)

**Input:** the molecular graph \( G \), the learned reverse model \( \epsilon_\theta \).

**Output:** the molecular conformation \( C \).

1: Sample \( C^T \sim p(C^T) = \mathcal{N}(0, I) \)
2: for \( s = T, T - 1, \ldots, 1 \) do
3: Shift \( C^s \) to zero CoM
4: Compute \( \mu_\theta(C^s, G, s) \) from \( \epsilon_\theta(C^s, G, s) \) using equation 4
5: Sample \( C^{s-1} \sim \mathcal{N}(C^s-1; \mu_\theta(C^s, G, s), \sigma^2 I) \)
6: end for
7: return \( C^0 \) as \( C \)
### Table 1: Results on the GEOM-Drugs dataset, without FF optimization.

<table>
<thead>
<tr>
<th>Models</th>
<th>COV-R (%) ↑</th>
<th>MAT-R (Å) ↓</th>
<th>COV-P (%) ↑</th>
<th>MAT-P (Å) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVGAE</td>
<td>0.00 0.00</td>
<td>3.0702 2.9937</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>GraphDG</td>
<td>8.27 0.00</td>
<td>1.9722 1.9845</td>
<td>2.08 0.00</td>
<td>2.4340 2.4100</td>
</tr>
<tr>
<td>CGCF</td>
<td>53.96 57.06</td>
<td>1.2487 1.2247</td>
<td>21.68 13.72</td>
<td>1.8571 1.8066</td>
</tr>
<tr>
<td>ConFDVAE</td>
<td>55.20 59.43</td>
<td>1.2380 1.1417</td>
<td>22.96 14.05</td>
<td>1.8287 1.8159</td>
</tr>
<tr>
<td>GeoMol</td>
<td>67.16 71.71</td>
<td>1.0875 1.0586</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>ConfigF</td>
<td>62.15 70.93</td>
<td>1.1629 1.1596</td>
<td>23.42 15.52</td>
<td>1.7219 1.6863</td>
</tr>
<tr>
<td><strong>GeoDiff-A</strong></td>
<td><strong>88.36 96.09</strong></td>
<td><strong>0.8704 0.8628</strong></td>
<td><strong>60.14 61.25</strong></td>
<td><strong>1.1864 1.1391</strong></td>
</tr>
<tr>
<td><strong>GeoDiff-C</strong></td>
<td><strong>89.13 97.88</strong></td>
<td><strong>0.8629 0.8529</strong></td>
<td><strong>61.47 64.55</strong></td>
<td><strong>1.1712 1.1232</strong></td>
</tr>
</tbody>
</table>

* The COV-R and MAT-R results of CVGAE, GraphDG, CGCF, and ConfigF are borrowed from Shi et al. (2021). The results of GeoMol are borrowed from a most recent study Zhu et al. (2022). Other results are obtained by our own experiments. The results of all models for the GEOM-QM9 dataset (summarized in Tab. 5) are collected in the same way.

Results & discussion. The results are summarized in Tab. 1 and Tab. 5 (left in Appendix. D). As noted in Sec. 4.3, GeoDiff can be trained with two types of modified ELBO, named alignment and chain-rule approaches. We denote models learned by these two objectives as GeoDiff-A and GeoDiff-C respectively. As shown in the tables, GeoDiff consistently outperform the state-of-the-art ML models on all datasets and metrics, especially by a significant margin for more challenging large molecules (Drugs dataset). The results demonstrate the superior capacity of GeoDiff to model the multi modal distribution, and generative both accurate and diverse conformations. We also notice that in general GeoDiff-C performs slightly better than GeoDiff-A, which suggests that chain-rule approach leads to a better optimization procedure. We thus take GeoDiff-C as the representative in the following comparisons. We visualize samples generated by different models in Fig. 2 to provide a qualitative comparison, where GeoDiff is shown to capture better both local and global structures.

On the more challenging Drugs dataset, we further test RDKit. As shown in Tab. 2, our observation is in line with previous studies (Shi et al., 2021) that the state-of-the-art ML models (shown in Tab. 1) perform better on COV-R and MAT-R. However, for the new Precision-based metrics we found that ML models are still not comparable. This indicates that ML models tend to explore more possible representatives while RDKit concentrates on a few most common ones, prioritizes quality over diversity. Previous works (Mansimov et al., 2019; Xu et al., 2021b) suggest that this is because RDKit involves an additional empirical force field (FF) (Halgren, 1996) to optimize the structure, and we follow them to also combine GeoDiff with FF to yield a more fair comparison. Results in
Figure 2: Examples of generated structures from Drugs dataset. For every model, we show the conformation best-aligned with the ground truth. More examples are provided in Appendix E.

<table>
<thead>
<tr>
<th>Models</th>
<th>COV-R (%)↑</th>
<th>MAT-R (Å)↓</th>
<th>COV-P (%)↑</th>
<th>MAT-P (Å)↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDKIT</td>
<td>Mean 60.91</td>
<td>Median  65.70</td>
<td>Mean 1.2026</td>
<td>Median 1.1252</td>
</tr>
<tr>
<td>GEOIFF + FF</td>
<td>Mean 92.27</td>
<td>Median 100.00</td>
<td>Mean 0.7618</td>
<td>Median 0.7340</td>
</tr>
</tbody>
</table>

Table 2: Results on the GEOM-Drugs dataset, with FF optimization.

Table 3: MAE of predicted ensemble properties in eV.

<table>
<thead>
<tr>
<th>Method</th>
<th>$E$</th>
<th>$E_{\text{min}}$</th>
<th>$\Delta \epsilon$</th>
<th>$\Delta \epsilon_{\text{min}}$</th>
<th>$\Delta \epsilon_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDKIT</td>
<td>0.9233</td>
<td>0.6585</td>
<td>0.3698</td>
<td>0.8021</td>
<td>0.2359</td>
</tr>
<tr>
<td>GRAPHDG</td>
<td>9.1027</td>
<td>0.8882</td>
<td>1.7973</td>
<td>4.1743</td>
<td>0.4776</td>
</tr>
<tr>
<td>CGCF</td>
<td>28.9661</td>
<td>2.8410</td>
<td>2.8356</td>
<td>10.6361</td>
<td>0.5954</td>
</tr>
<tr>
<td>CONFVAE</td>
<td>8.2080</td>
<td>0.6100</td>
<td>1.6080</td>
<td>3.9111</td>
<td>0.2429</td>
</tr>
<tr>
<td>CONFGF</td>
<td>2.7886</td>
<td>0.1765</td>
<td>0.4688</td>
<td>2.1843</td>
<td>0.1433</td>
</tr>
<tr>
<td>GEOIFF</td>
<td>0.25974</td>
<td>0.1551</td>
<td>0.3091</td>
<td>0.7033</td>
<td>0.1909</td>
</tr>
</tbody>
</table>

Tab. 2 demonstrate that GEOIFF +FF can keep the superior diversity (Recall metrics) while also enjoy significantly improved accuracy (Precision metrics).

5.3 Property Prediction

Evaluation metrics. This task estimates the molecular ensemble properties (Axelrod & Gomez-Bombarelli, 2020) over a set of generated conformations. This can provide an direct assessment on the quality of generated samples. In specific, we follow Shi et al. (2021) to extract a split from GEOM-QM9 covering 30 molecules, and generate 50 samples for each. Then we use the chemical toolkit PSI4 (Smith et al., 2020) to calculate each conformer’s energy $E$ and HOMOLUMO gap $\epsilon$, and compare the average energy $\overline{E}$, lowest energy $E_{\text{min}}$, average gap $\Delta \epsilon$, minimum gap $\Delta \epsilon_{\text{min}}$, and maximum gap $\Delta \epsilon_{\text{max}}$ with the ground truth.

Results & discussions. The mean absolute errors (MAE) between calculated properties and the ground truth are reported in Tab. 3. CVGAE is excluded due to the poor performance, which is also reported in Simm & Hernandez-Lobato (2020); Shi et al. (2021). The properties are highly sensitive to geometric structure, and thus the superior performance demonstrate that GEOIFF can consistently predict more accurate conformations across different molecules.

6 Conclusion

We propose GEOIFF, a novel probabilistic model for generating molecular conformations. GEOIFF marries denoising diffusion models with geometric representations, where we parameterize the reverse generative dynamics as a Markov chain, and novelty impose roto-translational invariance into the density with equivariant Markov kernels. We derive a tractable invariant objective from the variational lower bound to optimize the likelihood. Comprehensive experiments over multiple tasks demonstrate that GEOIFF is competitive with the existing state-of-the-art models. Future work includes further improving or accelerating the model with other recent progress of diffusion models, and extending our method to other challenging structures such as proteins.
ACKNOWLEDGEMENT

Minkai thanks Huiyu Cai, David Wipf, Zuobai Zhang, and Zhaocheng Zhu for their helpful discussions and comments. This project is supported by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, the Canada CIFAR AI Chair Program, collaboration grants between Microsoft Research and Mila, Samsung Electronics Co., Ltd., Amazon Faculty Research Award, Tencent AI Lab Rhino-Bird Gift Fund and a NRC Collaborative R&D Project (AI4D-CORE-06). This project was also partially funded by IVADO Fundamental Research Project grant PRF-2019-3583139727. The Stanford team is supported by NSF(#1651565, #1522054, #1733686), ONR (N000141912145), AFOSR (FA95501910024), ARO (W911NF-21-1-0125) and Sloan Fellowship.

REFERENCES


A PROOFS

A.1 PROPERTIES OF THE DIFFUSION MODEL

We include proofs for several key properties of the probabilistic diffusion model here to be self-contained. For more detailed discussions, please refer to Ho et al. (2020). Let \( \{\beta_t, \ldots, \beta_T\} \) be a sequence of variances, and \( \alpha_t = 1 - \beta_t \) and \( \bar{\alpha}_t = \prod_{s=1}^t \alpha_s \). The two following properties are crucial for deriving the final tractable objective in equation 9.

**Property 1.** Tractable marginal of the forward process:

\[
q(C^t | C^0) = \int q(C^{1:t} | C^0) \, dC^{1:t-1} = \mathcal{N}(C^t; \sqrt{\bar{\alpha}_t}C^0, (1 - \bar{\alpha}_t)I).
\]

*Proof.* Let \( \epsilon_t \)'s be independent standard Gaussian random variables. Then, by definition of the Markov kernels \( q(C^t | C^{t-1}) \) in equation 2, we have

\[
\begin{align*}
C^t &= \sqrt{\bar{\alpha}_t}C^{t-1} + \sqrt{\beta_t} \epsilon_t \\
&= \sqrt{\bar{\alpha}_t} \alpha_{t-1} C^{t-2} + \sqrt{\bar{\alpha}_t} \beta_{t-1} \epsilon_{t-1} + \sqrt{\beta_t} \epsilon_t \\
&= \sqrt{\bar{\alpha}_t} \alpha_{t-1} \alpha_{t-2} C^{t-3} + \sqrt{\bar{\alpha}_t} \alpha_{t-1} \beta_{t-2} \epsilon_{t-2} + \sqrt{\alpha_t} \beta_{t-1} \epsilon_{t-1} + \sqrt{\beta_t} \epsilon_t \\
&= \cdots \\
&= \sqrt{\bar{\alpha}_t} C^0 + \sqrt{\bar{\alpha}_t} \alpha_{t-1} \cdot \cdots \cdot \alpha_2 \beta_1 \epsilon_1 + \cdots + \sqrt{\alpha_t} \beta_{t-1} \epsilon_{t-1} + \sqrt{\beta_t} \epsilon_t.
\end{align*}
\]

Therefore \( q(C^t | C^0) \) is still Gaussian, and the mean of \( C^t \) is \( \sqrt{\bar{\alpha}_t} C^0 \), and the variance matrix is

\[
\begin{align*}
(\alpha_t \alpha_{t-1} \cdots \alpha_2 \beta_1 + \cdots + \alpha_t \beta_{t-1} + \beta_t)I = (1 - \bar{\alpha}_t)I.
\end{align*}
\]

Then we have:

\[
q(C^t | C^0) = \mathcal{N}(C^t; \sqrt{\bar{\alpha}_t}C^0, (1 - \bar{\alpha}_t)I).
\]

This property provides convenient closed-form evaluation of \( C^t \) knowing \( C^0 \):

\[
C^t = \sqrt{\bar{\alpha}_t} C^0 + \sqrt{\bar{\alpha}_t} \epsilon_t,
\]

where \( \epsilon_t \sim \mathcal{N}(0, I) \).

Besides, it is worth noting that,

\[
q(C^T | C^0) = \mathcal{N}(C^T; \sqrt{\bar{\alpha}_T}C^0, (1 - \bar{\alpha}_T)I),
\]

where \( \bar{\alpha}_T = \prod_{t=1}^T (1 - \beta_t) \) approaches zero with large \( T \), which indicates the diffusion process can finally converge into a whitened noisy distribution. \( \square \)

**Property 2.** Tractable posterior of the forward process:

\[
q(C^{t-1} | C^t, C^0) = \mathcal{N}(C^{t-1}; \frac{\alpha_{t-1} \beta_t}{1 - \bar{\alpha}_t} C^0 + \frac{\sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} C^t, (\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}) \beta_t I).
\]

*Proof.* Let \( \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \), then we can derive the posterior by Bayes rule:

\[
\begin{align*}
q(C^{t-1} | C^t, C^0) &= \frac{q(C^t | C^{t-1}) q(C^{t-1} | C^0)}{\mathcal{N}(C^t; \sqrt{\bar{\alpha}_t}C^0, (1 - \bar{\alpha}_t)I)} \\
&= \frac{(2\pi \beta_t)^{-\frac{d}{2}} (2\pi (1 - \bar{\alpha}_{t-1}))^{-\frac{d}{2}} (2\pi (1 - \bar{\alpha}_t))^{-\frac{d}{2}}}{2(1 - \bar{\alpha}_{t-1})} \mathcal{N}(C^{t-1}; \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t C^0, (\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}) \beta_t I) \\
&= \frac{\exp (-\frac{1}{2 \beta_t} \|C^{t-1} - \sqrt{\alpha_t} \beta_t C^0 - \frac{\sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} C^t \|^2)}{\bar{\alpha}_t I).
\end{align*}
\]

Then we have the posterior \( q(C^{t-1} | C^t, C^0) \) as the given form. \( \square \)
A.2 Proof of Proposition 1

Let $T_g$ be some roto-translation transformations of a group element $g \in \text{SE}(3)$, and let $p(x_T)$ be a density which is SE(3)-invariant, i.e., $p(x_T) = p(T_g(x_T))$. If the Markov transitions $p(x_{t-1}|x_t)$ are SE(3)-equivariant, i.e., $p(x_{t-1}|x_t) = p(T_g(x_{t-1})|T_g(x_t))$, then we have that the density $p_\theta(x_0) = \int p(x_T)p_\theta(x_{0:T-1}|x_T)dx_{1:T}$ is also SE(3)-invariant.

Proof.

\begin{equation}
\begin{aligned}
    p_\theta(T_g(x_0)) &= \int p(T_g(x_T))p_\theta(T_g(x_{0:T-1})|T_g(x_T))dx_{1:T} \\
    &= \int p(T_g(x_T))\Pi_{t=1}^T p_\theta(T_g(x_{t-1})|T_g(x_t))dx_{1:T} \\
    &= \int p(x_T)\Pi_{t=1}^T p_\theta(x_{t-1}|x_t)dx_{1:T} \quad \text{(invariant prior $p(x_T)$)} \\
    &= \int p(x_T)\Pi_{t=1}^T p_\theta(x_{t-1}|x_t)dx_{1:T} \quad \text{(equivariant kernels $p(x_{t-1}|x_t)$)} \\
    &= \int p(x_T)p_\theta(x_{0:T-1}|x_T)dx_{1:T} \\
    &= p_\theta(x_0)
\end{aligned}
\end{equation}

\$
\square
\$

A.3 Proof of Proposition 2

In this section we prove that the output $x$ of GFN defined in equation 5, 6 and 7 is translationally invariant and rotationally equivariant with the input $C$. Let $g \in \mathbb{R}^3$ denote any translation transformations and orthogonal matrices $R \in \mathbb{R}^{3\times 3}$ denote any rotation transformations. let $Rx$ be shorthand for $(Rx_1, \cdots, Rx_N)$. Formally, we aim to prove that the model satisfies:

$$Rx^{t+1}, h^{t+1} = \text{GFN}(Rx^t, RC + g, h^t).$$

(15)

This equation indicates that, given $x^t$ already rotationally equivalent with $C$, and $h^t$ already invariant, then such property can propagate through a single GFN layer to $x^{t+1}$ and $h^{t+1}$.

Proof. Firstly, given that $h^t$ already invariant to SE(3) transformations, we have that the messages $m_{ij}$ calculated from equation 5 will also be invariant. This is because it sorely relies on the distance between two atoms, which are manifestly invariant to rotations $\|Rx^t_i - Rx^t_j\|^2 = (x^t_i - x^t_j)^T R^T R (x^t_i - x^t_j)$, and let $\Phi_x$ be a shorthand for $(\Phi_x(x^t_i - x^t_j), \Phi_x(x^t_i - x^t_j))$. Formally, the invariance of messages in equation 5 can be written as:

$$m_{ij} = \Phi_m(h^t_i, h^t_j, \|Rx^t_i - Rx^t_j\|^2, e_{ij}) = \Phi_m(h^t_i, h^t_j, \|x^t_i - x^t_j\|^2, e_{ij}).$$

(16)

And similarly, the $h^{t+1}$ updated from equation 6 will also be invariant.

Next, we prove that the vector $x$ updated from equation 7 preserves rotational equivariance and translational invariance. Given $m_{ij}$ already invariant as proven above, we have that:

$$\sum_{j \in \mathcal{N}(i)} \frac{1}{d_{ij}} (Re_i + g - Re_j - g) \Phi_x(m_{ij}) = R \sum_{j \in \mathcal{N}(i)} \frac{1}{d_{ij}} (e_i - e_j) \Phi_x(m_{ij}) = Rx^{t+1}.$$ 

(17)

Therefore, we have that rotating and translating $c$ results in the same rotation and no translation on $x^{t+1}$ by updating through equation 7.

Thus we can conclude that the property defined in equation 15 is satisfied.

\$\square$
A.4 Proof of Proposition 3

We first derive the variational lower bound (ELBO) objective in equation 8. The ELBO can be calculated as follows:

$$
\mathbb{E} \log p_0(C^0|G) = \mathbb{E} \log \mathbb{E}_{q(C^1,T|C^0)} \left[ \frac{p_0(C^{0:T-1}|G, C^T) \times p(C^T)}{q(C^{1:T}|C^0)} \right] \\
\geq E_q \log \frac{p_0(C^{0:T-1}|G, C^T) \times p(C^T)}{q(C^{1:T}|C^0)} \\
= E_q \log p(C^T) - \sum_{t=1}^T \log \frac{p_0(C^{t-1}|G, C^t)}{q(C^t|C^{t-1})} \\
= E_q \left[ \log p(C^T) - \log \frac{p_0(C^0|G, C^1)}{q(C^1|C^0)} - \sum_{t=2}^T \left( \log \frac{p_0(C^{t-1}|G, C^t)}{q(C^t|C^{t-1})} + \log \frac{q(C^{t-1}|C^0)}{q(C^t|C^0)} \right) \right] \\
= E_q \left[ \log \frac{p(C^T)}{q(C^T|C^0)} - \log p_0(C^0|G, C^1) - \sum_{t=2}^T \log \frac{p_0(C^{t-1}|G, C^t)}{q(C^t|C^{t-1})} \right] \\
= -E_q \left[ KL(q(C^T|C^0)||p(C^T)) + \sum_{t=2}^T KL(q(C^{t-1}|C^t, C^0)||p_0(C^{t-1}|G, C^t)) - \log p_0(C^0|G, C^1) \right].
$$

(18)

It can be noted that the first term $KL(q(C^T|C^0)||p(C^T))$ is a constant, which can be omitted in the objective. Furthermore, for brevity, we also merge the final term $\log p_0(C^0|G, C^1)$ into the second term (sum over KL divergences), and finally derive that $\mathcal{L}_{ELBO} = \sum_{t=1}^T D_{KL}(q(C^{t-1}|C^t, C^0)||p_0(C^{t-1}|G, C^t))$ as in equation 8.

Now we consider how to compute the KL divergences as the proposition 3. Since both $q(C^{t-1}|C^t, C^0)$ and $p_0(C^{t-1}|G, C^t)$ are Gaussian share the same covariance matrix $\beta_t I$, the KL divergence between them can be calculated by the squared $\ell_2$ distance between their means weighed by a certain weights $\frac{1}{2\beta_t}$. By the expression of $q(C^t|C^0)$, we have the reparameterization that $C^t = \sqrt{\alpha_t}C^0 + \sqrt{1-\alpha_t}\epsilon$. Then we can derive:

$$
\mathbb{E}_t KL(q(C^{t-1}|C^t, C^0)||p_0(G, C^{t-1}|C^t)) = \frac{1}{2\beta_t} \mathbb{E} \left\| \sqrt{\alpha_t - \beta_t} \bar{C}^0 + \sqrt{\alpha_t(1 - \alpha_t - \beta_t)} C^t - \frac{1}{\sqrt{\alpha_t}} \left( C^t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon \theta(C^t, G, t) \right) \right\|^2 \\
= \frac{1}{2\beta_t} \mathbb{E} \left\| \sqrt{\alpha_t - \beta_t} \bar{C}^0 + \sqrt{\alpha_t(1 - \alpha_t - \beta_t)} C^t - \frac{1}{\sqrt{\alpha_t}} \left( C^t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon \theta(C^t, G, t) \right) \right\|^2 \\
= \frac{1}{2\beta_t} \mathbb{E} \left\| \sqrt{\alpha_t(1 - \alpha_t)} C^0, \epsilon \left\| 0 \cdot C^t + \epsilon - \epsilon \theta(C^t, G, t) \right\| \right\|^2 \\
= \frac{1}{2\beta_t} \mathbb{E} \left\| \epsilon \left\| \epsilon - \epsilon \theta(C^t, G, t) \right\| \right\|^2 \\
= \gamma_t \mathbb{E} \left\| \epsilon \right\|^2,
$$

(19)

where $\gamma_t$ represent the weights $\frac{\beta_t}{2\alpha_t(1-\alpha_t)}$. And we finish the proof.

A.5 Analysis of the invariant density in Sec. 4.2

Given a geometric system $x \in \mathbb{R}^{N,3}$, we obtain the CoM-free $\hat{x}$ by subtracting its CoM. This can be considered as a linear transformation:

$$
\hat{x} = Q x, \quad \text{where} \quad Q = I_3 \otimes \left( I_N - \frac{1}{N} 1_N 1_N^T \right)
$$

(20)

where $I_k$ denotes the $k \times k$ identity matrix and $1_k$ denotes the $k$-dimensional vector filled with ones. It can be noted that $Q$ is a symmetric projection operator, i.e., $Q^2 = Q$ and $Q^T = Q$. And we also
have that \( \text{rank}(Q) = (N - 1) \cdot 3 \). Furthermore, let \( U \) represent the space of CoM-free systems, we can easily have that \( Qy = y \) for any \( y \in U \) since the CoM of \( y \) is already zero.

Formally, let \( u = N \cdot 3 \) and set \( \mathbb{R}^u \) with an isotropic normal distribution \( \rho = N(0, I_u) \), then the CoM-free density can be formally written as \( \hat{\rho} = N(0, QQ^T) = N(0, Q\mathbb{I}Q^T) \). Thus, sampling from \( \rho \) can be trivially achieved by sampling from \( \hat{\rho} \) and then projecting with \( Q \). And \( \hat{\rho}(y) \) can be calculated by \( \rho(y) \) since for any \( y \in U \) we have \( \|y\|_2^2 = \|Qy\|_2^2 \), and thus \( \rho(y) = \hat{\rho}(y) \).

And in this paper, with the SE(3)-equivariant Markov kernels of the reverse process, any CoM-free system will transit to another CoM-free system. And thus we can induce a well-defined Markov chain on the subspace spanned by \( Q \).

## B OTHER RELATED WORK

**Protein structure generation.** There has also been many recent works working on protein structure folding. For example, Boltzmann generators Noé et al. (2019) use flow-based models to generate the structure of protein main chains. AlQuraishi (2019) uses recurrent networks to model the amino acid sequences. Ingraham et al. (2019) proposed neural networks to learn an energy simulator to infer the protein structures. Most recently, AlphaFold Senior et al. (2020); Jumper et al. (2021) has significantly improved the performance of protein structure generation. Nevertheless, proteins are mainly linear backbone structures while general molecules are highly branched with various rings, making protein folding approaches unsuitable for our setting.

**Point cloud generation.** Recently, some other works (Luo & Hu, 2021; Chibane et al., 2020) has also been proposed for 3D structure generation with diffusion-based models, but focus on the point cloud problem. Unfortunately, in general, point clouds are not considered as graphs with various atom and bond information, and equivariance is also not widely considered, making these methods fundamentally different from our model.

## C EXPERIMENT DETAILS

In this section, we introduce the details of our experiments. In practice, the means \( \epsilon_\theta \) are parameterized as compositions of both typical invariant MPNNs (Schütt et al., 2017) and the proposed equivariant GFNs in Sec. 4.2. As a default setup, the MPNNs for parameterizing the means \( \epsilon_\theta \) are all implemented with 4 layers, and the hidden embedding dimension is set as 128. After the MPNNs, we can obtain the informative invariant atom embeddings, which we denote as \( h_0 \). Then the embeddings \( h_0 \) are fed into equivariant layers and updated with equation 5, equation 6, and equation 7 to obtain the equivariant output. For the training of GEODIFF, we train the model on a single Tesla V100 GPU with a learning rate of 0.001 until convergence and Adam (Kingma & Welling, 2013) as the optimizer. The practical training time is \( \sim \)48 hours. The other hyper-parameters of GEODIFF are summarized in Tab. 4, including highest variance level \( \beta_T \), lowest variance level \( \beta_T \), the variance schedule, number of diffusion timesteps \( T \), radius threshold for determining the neighbor of atoms \( \tau \), batch size, and number of training iterations.

<table>
<thead>
<tr>
<th>Task</th>
<th>( \beta_1 )</th>
<th>( \beta_T )</th>
<th>( \beta ) scheduler</th>
<th>( T )</th>
<th>( \tau )</th>
<th>Batch Size</th>
<th>Train Iter.</th>
</tr>
</thead>
<tbody>
<tr>
<td>QM9</td>
<td>1e-7</td>
<td>2e-3</td>
<td>sigmoid</td>
<td>5000</td>
<td>10Å</td>
<td>64</td>
<td>1M</td>
</tr>
<tr>
<td>Drugs</td>
<td>1e-7</td>
<td>2e-3</td>
<td>sigmoid</td>
<td>5000</td>
<td>10Å</td>
<td>32</td>
<td>1M</td>
</tr>
</tbody>
</table>

## D ADDITIONAL EXPERIMENTS

### D.1 RESULTS FOR GEOM-QM9

The results on the GEOM-QM9 dataset are reported in Tab. 5.
Table 5: Results on the GEOM-QM9 dataset, without FF optimization.

<table>
<thead>
<tr>
<th>Models</th>
<th>COV-R (%) ↑</th>
<th>MAT-R (Å) ↓</th>
<th>COV-P (%) ↑</th>
<th>MAT-P (Å) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>CVGAE</td>
<td>0.09</td>
<td>0.00</td>
<td>1.6713</td>
<td>1.6088</td>
</tr>
<tr>
<td>GRAPHDG</td>
<td>73.33</td>
<td>84.21</td>
<td>0.4245</td>
<td>0.3973</td>
</tr>
<tr>
<td>CGCF</td>
<td>78.05</td>
<td>82.48</td>
<td>0.4219</td>
<td>0.3900</td>
</tr>
<tr>
<td>CONFVAE</td>
<td>77.84</td>
<td>88.20</td>
<td>0.4154</td>
<td>0.3739</td>
</tr>
<tr>
<td>GeoMOL</td>
<td>71.26</td>
<td>72.00</td>
<td>0.3731</td>
<td>0.3731</td>
</tr>
<tr>
<td>CONFGF</td>
<td>88.49</td>
<td>94.31</td>
<td>0.2673</td>
<td>0.2685</td>
</tr>
<tr>
<td>GeoDiff-A</td>
<td>90.54</td>
<td>94.61</td>
<td>0.2104</td>
<td>0.2021</td>
</tr>
<tr>
<td>GeoDiff-C</td>
<td>90.07</td>
<td>93.39</td>
<td>0.2090</td>
<td>0.1988</td>
</tr>
</tbody>
</table>

Table 6: Additional results on the GEOM-Drugs dataset, without FF optimization.

<table>
<thead>
<tr>
<th>Models</th>
<th>COV-R (%) ↑</th>
<th>MAT-R (Å) ↓</th>
<th>COV-P (%) ↑</th>
<th>MAT-P (Å) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>GeoDiff (T=1000)</td>
<td>82.96</td>
<td>96.29</td>
<td>0.9525</td>
<td>0.9334</td>
</tr>
</tbody>
</table>

D.2 Ablation study with fewer diffusion steps

We also test our method with fewer diffusion steps. Specifically, we test the setting with $T = 1000$, $\beta_1 = 1e-7$ and $\beta_T = 9e-3$. The results on the more challenging Drugs dataset are shown in Tab. 6. Compared with the results in Tab. 1, we can observe that when setting the diffusion steps as 1000, though slightly weaker than the performance with 5000 decoding steps, the model can already outperforms all existing baselines. Note that, the most competitive baseline CONFGF (Shi et al., 2021) also requires 5000 sampling steps, which indicates that our model can achieve better performance with fewer computational costs compared with the state-of-the-art method.

E More Visualizations

We provide more visualization of generated structures in Fig. 3. The molecules are chosen from the test split of GEOM-Drugs dataset.
<table>
<thead>
<tr>
<th>Graph</th>
<th>Conformations</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Graph 1" /></td>
<td><img src="image2.png" alt="Conformations 1" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Graph 2" /></td>
<td><img src="image4.png" alt="Conformations 2" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Graph 3" /></td>
<td><img src="image6.png" alt="Conformations 3" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Graph 4" /></td>
<td><img src="image8.png" alt="Conformations 4" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Graph 5" /></td>
<td><img src="image10.png" alt="Conformations 5" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Graph 6" /></td>
<td><img src="image12.png" alt="Conformations 6" /></td>
</tr>
</tbody>
</table>

Figure 3: Visualization of drug-like conformations generated by GEODIFF.