Non-Differentiable Diffusion Guidance for Improved Molecular Geometry

Yuchen Shen [*1] **Chenhao Zhang** [*2] **Chenghui Zhou** [*2] **Sijie Fu** [3] **Newell Washburn** [3 4] **Barnabás Póczos** [2]

Abstract

Diffusion models are a promising approach to generate molecules in 3D. However, challenges remain in generating realistic 3D molecules with refined geometry that respect the quantum physical laws governing the atom configurations. In this work, we generalized the neural predictor used in diffusion guidance to a *non-differentiable* expert oracle, GFN2-xTB, a semi-empirical quantum mechanical method for accurate and efficient quantum chemistry calculations. With an off-the-shelf diffusion model, we guide it to generate molecules that are valid and more stable with less net force. By prompting atoms to move to lower molecular energy with the estimated gradient from GFN2-xTB, we show that our method generates molecules that are more stable and favored in energy with better-optimized geometries than existing literature.

1. Introduction

Applications of generative models in feature-rich geometries have the potential to accelerate scientific discoveries in chemistry, biology, and materials science. For example, the *in silico* generation of 3D geometries for molecules and proteins can help screen novel drug candidates and model the protein-molecule interaction to accelerate drug discovery (Corso et al., 2022; Jumper et al., 2021; Xu et al., 2023; Hoogeboom et al., 2022). Particularly, molecules can be modeled with a graph, where each node of the graph represents an atom and contains feature information such as 3D coordinates and atom types. The geometries of the generation results have domain-specific implications – a molecule's stability and properties depend significantly on its preferred quantum geometric states, i.e., atomic and molecular geometries. For example, polarity is closely related to molecular geometry. A water molecule H_2O has a stable V-shaped H-O-H geometry of 104.5 degrees and is thus polar. A generated H_2O molecule with a linear H-O-H geometry would instead be nonpolar but unstable. Therefore, when we discuss molecules and their properties, it is essential to start from their preferred stable geometries. In this paper, we propose to improve generated molecular geometries by incorporating a non-differentiable oracle in the diffusion sampling process.

Many diffusion model approaches have been applied to 3D molecular structures (Hoogeboom et al., 2022; Xu et al., 2023; Bao et al., 2022; Vignac et al., 2022). The generated molecules are evaluated on general stability and validity as well as on the performance of property-conditioned generation, all of which depend on the geometry of the resulting molecules. To better control the generation results, Hoogeboom et al. (2022); Xu et al. (2023) have proposed to train *conditional* generative models, where they input the property values as conditions into the model during training and sampling to obtain novel molecules fulfilling the requirement. Inspired by guidance algorithms that improve image conditional generation quality (Bansal et al., 2023; Dhariwal & Nichol, 2021), Han et al. (2023) applied the training-free regressor guidance on the sampling process of an *unconditional* diffusion model, and have seen remarkable improvements for molecule conditional generation. Compared with training a conditional diffusion model from scratch, the guidance method achieves conditional generation by steering an unconditional model with a neural classifier or regressor, which requires fewer computation resources to train. However, it requires computing the gradient through the *differentiable* neural predictor to perform "correction" during the sampling process, which hinders its broader applications as neural networks tend to suffer from extrapolation, and many chemical properties are calculated by *non-differentiable* oracles that can't be backpropagated.

To address the above issues, we propose to generalize the use of a neural predictor in diffusion guidance to a *non-differentiable* expert oracle which we can query from. To improve the generated molecules' geometries, we use an external quantum chemistry package *xTB* with the GFN2-xTB method (Bannwarth et al., 2019) that conducts accurate and efficient quantum chemistry calculation for atom forces,

[*]Equal contribution [1]Language Technologies Institute, Carnegie Mellon University [2]Machine Learning Department, Carnegie Mellon University [3]Department of Chemistry, Carnegie Mellon University [4]Department of Biomedical Engineering, Carnegie Mellon University. Correspondence to: Barnabás Póczos <bapoczos@cs.cmu.edu>.

ICML 2024 AI for Science workshop.

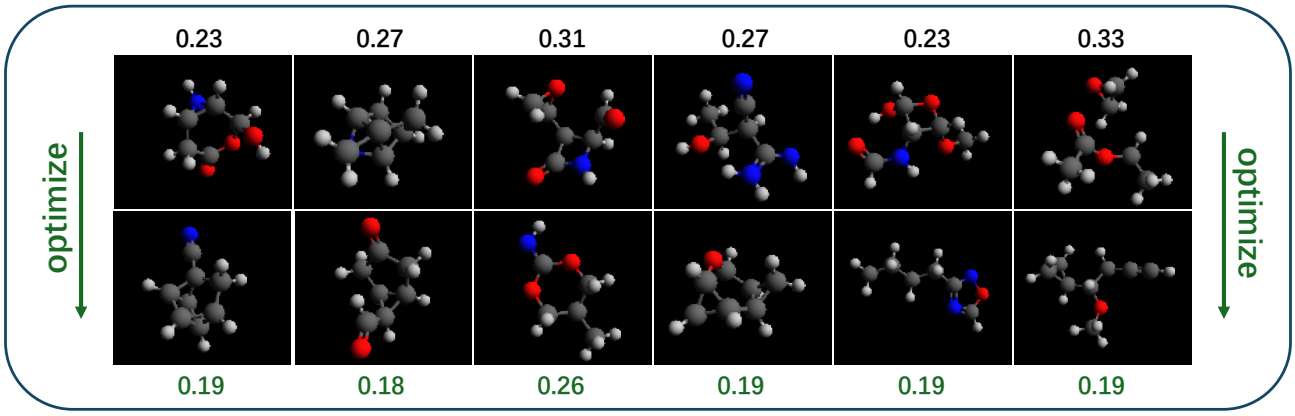

Figure 1. 3D visualization of molecules before (top row) and after (bottom row) geometry optimization using unconditional GeoLDM with non-differentiable xTB guidance. The numbers represent the L-1 norm of the atom forces calculated by xTB, where lower forces indicate a more stable structure. Note that the pre-optimized molecule at rightmost is broken into two parts, but our optimization method removes the disconnection and lowers the forces from 0.33 to 0.19.

and we use a two-point method (Nesterov & Spokoiny, 2017) to estimate the gradient of the generated molecule in the chemical property value landscape. By guiding the sampling process towards an output that minimizes forces on each atom, we naturally achieve geometry refinement as a result. Directly incorporating an expert oracle removes the estimation error of a predictor, because even in the unlikely event that the predictor achieves perfect accuracy on the test set, it is not guaranteed to achieve good accuracy on an unseen dataset with a different distribution. For example, the molecules from QM9 (Ramakrishnan et al.) are stable molecules with optimal energies and minimized force on each atom, which is usually not the truth for the molecules generated during the denoising process of a diffusion model; thus, a zero-error predictor on QM9 will not necessarily yield accurate predictions for the denoising molecules.

In this paper, we develop a non-differentiable diffusion guidance method and apply it to molecule generation for geometry optimization. To evaluate the effectiveness of our proposed approach, we show that our non-differentiable oracle guidance can improve validity and stability among the generated molecules in comparison with unconditional diffusion models. We present some sampled molecules with optimized geometry in Figure 1.

2. Related Work

This paper lies in the intersection of predictor-guided diffusion generation and molecule generation for drug discovery. Various types of generative models have been proposed to model molecular data, including VAE (Jin et al., 2018; 2019; Kusner et al., 2017; Maziarz et al., 2021) and GAN (Prykhodko et al., 2019; De Cao & Kipf, 2018). However, these models focus on generating molecules as 2D graphs without 3D coordinate information. The autoregressive model is one approach to generate 3D molecules including G-Schnet (Gebauer et al., 2019) and G-SphereNet (Luo & Ji, 2022), however, they are less effective and powerful compared to diffusion models (Hoogeboom et al., 2022). The equivariant diffusion model EDM for 3D molecule generation was first proposed by Hoogeboom et al. (2022), which utilizes an equivariant graph neural network to model the molecules as graphs with coordinates and atom types as node features. GeoLDM (Xu et al., 2023) further extends EDM to a latent diffusion architecture and has shown improvements in stability and validity. However, to achieve conditional generation, both EDM and GeoLDM need to be re-trained, where the target property value is appended to the feature space to generate molecules that fulfill certain property requirements.

Guided diffusion generation has shown promising results in the domain of image (Dhariwal & Nichol, 2021; Bansal et al., 2023; Zhang et al., 2023; Rombach et al., 2022), where its generation can be conditioned on texts (Rombach et al., 2022), poses and edges (Zhang et al., 2023), and classifiers (Dhariwal & Nichol, 2021). A similar approach is adopted for property-guided molecule generation (Vignac et al., 2022; Bao et al., 2022; Han et al., 2023). Without re-training a conditional model from scratch, Vignac et al. (2022); Han et al. (2023) trained additional differentiable neural regressors of properties, and use the gradients as guidance during the sampling process of an unconditional diffusion model. Bao et al. (2022) proposed to train time-dependent regressors of properties as guidance during sampling. Our paper differs from the previous work by introducing an approach to directly estimate the guidance signal (e.g., the gradient) from a non-differentiable expert oracle without training additional neural predictors.

3. Preliminary

In this section, we will introduce the diffusion model, the architecture we use and how it achieves equivariance, and the semi-empirical quantum mechanic method GFN2-xTB used as guidance in our paper.

Diffusion Models In general, a diffusion model (Ho et al., 2020; Song et al., 2020; Dhariwal & Nichol, 2021; Sohl-Dickstein et al., 2015) consists of a forward *diffusion process* and a reverse *denoising process*. The diffusion process is a Markov chain that gradually adds Gaussian noises with a variance schedule $\beta_{1:T}$ from timestep 1 to T to the original datapoint $\boldsymbol{x}_0$. The schedule is chosen such that $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The forward diffusion process q is usually defined as a fixed schedule by the following:

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$$
$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

where $\beta_{1:T}$ is pre-defined and fixed. The reverse denoising process starts with $\boldsymbol{x}_T$ and recovers the original datapoint $\boldsymbol{x}_0$ by predicting the mean of the distribution of $\boldsymbol{x}_{t-1}$ given $\boldsymbol{x}_t$, denoted as $\mu_\theta(\boldsymbol{x}_t, t)$ where θ is the parameter. We can model the reverse process as:

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$$
$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)) \quad (2)$$

In practice, Σ_θ is set to be $\sigma_t^2 \mathbf{I}$ for all t for simplicity, where $\sigma_t = \sqrt{1-\alpha_t^2}$ and $\alpha_t = \sqrt{\prod_{i=1}^{t}(1-\beta_i)}$. Ho et al. (2020) further simplified the objective from predicting mean $\mu_\theta(\boldsymbol{x}_t, t)$ to predict the noise at each step, which is denoted by $\epsilon_\theta(\boldsymbol{x}_t, t)$, as $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \epsilon$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The training objective is minimize $\mathbb{E}_{\boldsymbol{x}_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t)\|^2\right]$, and $\mu_\theta(\boldsymbol{x}_t, t)$ can be parameterized as $\frac{1}{1-\beta_t}(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t^2}}\epsilon_\theta(\boldsymbol{x}_t, t))$. Consequently, we have

$$\boldsymbol{x}_{t-1} \sim \mathcal{N}(\frac{1}{1-\beta_t}(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t^2}}\epsilon_\theta(\boldsymbol{x}_t, t)), \sigma_t^2 \mathbf{I}) \quad (3)$$

Latent Diffusion Architecture for 3D Molecule Generation An N-atom molecule can be represented as a point cloud $\mathcal{G} = [\mathbf{x}, \mathbf{h}] \in \mathbb{R}^{N \times (3+d)}$, with $\mathbf{x} \in \mathbb{R}^{N \times 3}$ being the N atom's 3D coordinates and $\mathbf{h} \in \mathbb{R}^{N \times d}$ as the atom features such as atoms types and charges. A latent diffusion architecture (Rombach et al., 2022; Xu et al., 2023) consists of a VAE and a diffusion model, and are trained consecutively. Particularly, the geometric latent diffusion model (GeoLDM) (Xu et al., 2023) uses the encoder of the VAE to project discrete molecules to a continuous latent space, on which the diffusion model is then trained. Denote the encoder as $\mathcal{E}$ and the latent variable by $\mathbf{z} \in \mathbb{R}^{N \times (3+d_z)}$, then $[\mathbf{z}_{\mathbf{x},0}, \mathbf{z}_{\mathbf{h},0}] = \mathcal{E}([\mathbf{x}, \mathbf{h}])$, with $\mathbf{z}_{\mathbf{h},0} \in \mathbb{R}^{N \times d_z}$ and $d_z < d$. Let $\mathbf{z}_t = [\mathbf{z}_{\mathbf{x},t}, \mathbf{z}_{\mathbf{h},t}]$, the latent forward diffusion process and reverse denoising process are defined as:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}) \quad (4)$$
$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)) \quad (5)$$

We denote the decoder of the VAE as $\mathcal{D}$, which maps $\mathbf{z}_0$ back to the original molecular space, such that $\mathcal{D}([\mathbf{z}_{\mathbf{x},0}, \mathbf{z}_{\mathbf{h},0}]) = [\mathbf{x}, \mathbf{h}] \in \mathbb{R}^{N \times (3+d)}$.

The encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ are parameterized with an equivariant graph neural network (EGNN) (Satorras et al., 2021) to translate between discrete molecular data and latent variables, such that the atom types are invariant and the positions are equivariant to transformations as follows:

$$R\mathbf{z}_{\mathbf{x},t} + T, \mathbf{z}_{\mathbf{h},t} = \mathcal{E}(R\mathbf{x}_t + T, \mathbf{h}_t)$$
$$R\mathbf{x}_t + T, \mathbf{h}_t = \mathcal{D}(R\mathbf{z}_{\mathbf{x},t} + T, \mathbf{z}_{\mathbf{h},t}) \quad (6)$$

for any rotation matrix R and translation matrix T, where $\mathbf{z}_{\mathbf{x},t} \in \mathbb{R}^{N \times 3}$ are required to satisfy zero center gravity and have zero-mean over N atoms for each position. In addition, the latent diffusion model is also parameterized by EGNN such that transitions between each timestep in the denoising process also respect the same characteristics.

GFN2-xTB Method According to the laws of physics and thermodynamics, matter such as electrons, atoms, and molecules, interacts with matter inherently to reach configurations with lower potential energies for better stability. To formulate this as a molecular geometry optimization problem, let $h_1, \cdots, h_N$ be the N atoms in a given molecule and $\mathbf{x_1}, \cdots, \mathbf{x_N} \in \mathbb{R}^3$ be their corresponding coordinates in the 3D space, then each atom h_i is subject to a force

$$\mathbf{f_i}(\mathcal{G}) = \frac{\partial E_p\left(\mathbf{x_1}, \cdots, \mathbf{x_N} \mid h_1, \cdots, h_N\right)}{\partial \mathbf{x_i}}, \forall i \in [N] \quad (7)$$

where E_p represents the potential energy of the conformation. The force here manifests valid physical interpretations: an atom is pushed accordingly by the exerted force until the force reduces to zero and an equilibrium is achieved. One necessary (but not sufficient) condition for a stable molecular geometry is that all forces on the atoms should be (close to) zero, i.e., $\forall i \in [N], \mathbf{f_i}(\mathcal{G}) = 0$.

However, the exact mathematical potential energy evaluation, i.e., the solution to the Schrödinger equation, is still a black box to us (Cao et al.). Over the years, different levels of theories and methods have been developed to evaluate the potential energy, such as force field (FF), semi-empirical methods (e.g., xTB), and density functional theory (DFT)

methods (e.g., B3LYP/6-31G(2df,p)). The methods are listed in order of increased accuracy and cost. After trading-off between accuracy and efficiency within a feasible computation cost, we selected GFN2-xTB, a more recent and advanced semi-empirical method (Bannwarth et al., 2019), to calculate the forces of the generated molecular geometry in the diffusion process. More details about the GFN2-xTB method can be found in Appendix B.

4. Methodology

Recall that molecular data is represented in the form of a point cloud $\mathcal{G} = [\mathbf{x}, \mathbf{h}] \in \mathbb{R}^{N\times(3+d)}$, we denote the generated molecule by the diffusion model as $\mathcal{G}_0$. We introduce a training-free guided denoising process by incorporating the GFN2-xTB method at inference time to minimize the net force acting on each atom and optimize the molecular geometry towards a more stable atom configuration. We denote the molecule generated by our method as $\mathcal{G}_0'$ and we aim to achieve $\frac{1}{N}\sum_i^N \mathbf{f_i}(\mathcal{G}_0') < \frac{1}{N}\sum_i^N \mathbf{f_i}(\mathcal{G}_0)$ for $i \in [N]$, with $\mathbf{f_i}(\mathcal{G})$ defined in Eq. 7.

Up next, we will first describe how to obtain guidance from a differentiable neural network (NN) (Vignac et al., 2022; Han et al., 2023; Bao et al., 2022), then introduce how to use a non-differentiable oracle in place of a differentiable NN to obtain the gradient for diffusion guidance.

4.1. Guidance for Diffusion Models

The goal of neural guidance is to direct the denoising process towards a target property value y. Dhariwal & Nichol (2021) proposed a method to modify the denoising process to achieve conditional generation with an unconditional diffusion model, with a scalar s controlling the guidance strength:

$$x_{t-1} \sim \mathcal{N}(\frac{1}{1-\beta_t}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t^2}}\epsilon_\theta(x_t, t)) + s\sigma_t^2\nabla_{x_t}\log p_\phi(y|x_t), \sigma_t^2\mathbf{I}) \tag{8}$$

In their case, $p_\phi(y|\boldsymbol{x}_t)$ is parameterized by a classifier and y is a categorical label, where the additional term $s\sigma_t^2\nabla_{\boldsymbol{x}_t}\log p_\phi(y|\boldsymbol{x}_t)$ shifts the mean of the sampling distribution to provide guidance. However, we are interested in the case of $y \in \mathbb{R}$, which is continuous. Let $f_\eta : \mathcal{G} \to \mathbb{R}$ be a NN that predicts the property score, where we follow Vignac et al. (2022) and assume $p(y|\boldsymbol{x}_t) \sim \mathcal{N}(f_\eta(\boldsymbol{x}_t), \sigma_\eta^2\mathbf{I})$ and $\sigma_\eta^2 = 1$ for simplicity, we can estimate that:

$$\begin{aligned}\nabla_{\boldsymbol{x}_t}\log p_\phi(y \mid \boldsymbol{x}_t) &\propto -\nabla_{\boldsymbol{x}_t}\|y - f_\eta(\boldsymbol{x}_t)\|_2^2 \\ &= -\nabla_{\boldsymbol{x}_t}\mathcal{L}(y, f_\eta(\boldsymbol{x}_t))\end{aligned} \tag{9}$$

where $\mathcal{L}(y, f_\eta(\boldsymbol{x}_t))$ is the MSE between the target and predicted value by $f_\eta(\cdot)$.

However, in the early stage of the denoising process, $\boldsymbol{x}_t = \alpha_t\boldsymbol{x}_0 + \sigma_t\epsilon$ might not be informative enough to predict y as it consists mostly of Gaussian noise. For more effective prediction during the denoising process, we can estimate the denoised version of $\boldsymbol{x}_t$ as (Kawar et al., 2022; Song et al., 2020):

$$\hat{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{x}_t, t)}{\sqrt{\alpha_t}} \tag{10}$$

And $\hat{\boldsymbol{x}}_0$ can be used in place of $\boldsymbol{x}_t$ in Eq (9) when calculating the guidance, such that $\nabla_{\boldsymbol{x}_t}\log p_\phi(y|\boldsymbol{x}_t) \approx \nabla_{\boldsymbol{x}_t}\log p_\phi(y|\hat{\boldsymbol{x}}_0)$.

4.2. Guidance From a Non-Differentiable Oracle

In this section, we aim to tackle a more challenging problem where the guidance is specified by a non-differentiable oracle. Since we adopted the latent diffusion architecture such that the diffusion model is trained on the continuous latent space encoded by VAE, we change our notation $\boldsymbol{x}_t$ for each state in the diffusion model in Section 4.1 to $\mathbf{z}_t$ as in Section 3. To precisely formulate our problem, which aims to refine the geometry of the generated molecules by minimizing the net force acting on each atom, our non-differentiable function $\mathbf{f}$ for guidance can be defined as the following:

$$\mathbf{f}(\mathcal{G}) = \frac{1}{N}\sum_i^N \mathbf{f_i}(\mathcal{G}) \tag{11}$$

where the molecular graph $\mathcal{G}$ is obtained by decoding the continuous latent variable through the decoder $\mathcal{G} = \mathcal{D}(\hat{\mathbf{z}}_0)$, $\hat{\mathbf{z}}_0$ is estimated by Eq. 10 and $\mathbf{f_i}$ is our non-differentiable oracle defined in Eq. 7 that yields the net force on each atom i, and our target value y in this case is 0.

When $\mathbf{f}$ is non-differentiable, we can no longer plug it into Eq. 9 to obtain the gradient as guidance. Instead, we can compute the gradient analytically. For ease of discussion, we define the one-step estimation of the original datapoint $\mathbf{z}_0$, introduced in Eq 10, as the function $t_0(\cdot)$:

$$\hat{\mathbf{z}}_0 = t_0(\mathbf{z}_t) = \frac{\mathbf{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{z}_t, t)}{\sqrt{\alpha_t}} \tag{12}$$

Recall that $\mathcal{L}(y, \mathbf{f}(\mathcal{G}))$ is the MSE loss function, $\mathbf{f} : \mathcal{G} \to \mathbb{R}$ is the non-differentiable oracle, and $\mathcal{D}$ is the decoder. We denote $\mathcal{F}$ as the composition $\mathbf{f} \circ \mathcal{D} \circ t_0$, the gradient is measured analytically by

$$\begin{aligned}\hat{\nabla}_{\mathbf{z}_t}\log p_\phi(y|\mathbf{z}_t) &\propto -\nabla_{\mathcal{F}(\mathbf{z}_t)}\mathcal{L}(y, \mathcal{F}(\mathbf{z}_t))\nabla_{\mathbf{z}_t}\mathcal{F}(\mathbf{z}_t) \\ &\approx -\nabla_{\mathcal{F}(\mathbf{z}_t)}\mathcal{L}(y, \mathcal{F}(\mathbf{z}_t)) \\ &\quad \cdot \lim_{\zeta\to 0}\frac{\mathcal{F}(\mathbf{z}_{\mathbf{x},t}\hat{\oplus}\zeta) - \mathcal{F}(\mathbf{z}_{\mathbf{x},t}\hat{\oplus}(-\zeta))}{2\zeta}\end{aligned} \tag{13}$$

where we define $\mathbf{z}_t \hat{\oplus} \zeta \coloneqq [\mathbf{z}_{\mathbf{x},t} + \zeta \mathbf{I}_{N\times 3}, \mathbf{z}_{\mathbf{h},t}]$. This approximation is possible because $\mathbf{z}_t$ is continuous after the projection of the VAE encoder. The formulation allows for guidance from a non-differentiable oracle such as quantum chemical method GFN2-xTB (Bannwarth et al., 2019). Estimating the gradient directly from an expert oracle eliminates the need to train additional neural regressors for properties such as the force on each atom, which comes with approximation error (Gasteiger et al., 2020). In addition, these regressors are often trained on stable molecules from QM9 instead of the potentially unstable molecules as byproducts of the denoising process, making them dubious options for property guidance.

However, directly adding $\pm\zeta \mathbf{1}_{N\times 3}$ to $\mathbf{z}_{\mathbf{x},t}$ would break the equivariance requirement in Eq. 6 on the latent variables, as it shifts the mean of the coordinates by ζ. To maintain zero center gravity of input to the $\mathcal{F}(\cdot)$, we construct a perturbation matrix $\mathbf{U} \in \mathbb{R}^{N\times 3}$ where each element is sampled from $\mathcal{N}(0,1)$, and apply Simultaneous Perturbation Stochastic Approximation (SPSA) (Spall, 1992; Nesterov & Spokoiny, 2017; Malladi et al., 2023) to estimate the gradient; thus, Eq. 13 becomes

$$\hat{\nabla}_{\mathbf{z}_t} \log p_\phi(y|\mathbf{z}_t) \propto -\nabla_{\mathcal{F}(\mathbf{z}_t)} \mathcal{L}(y, \mathcal{F}(\mathbf{z}_t)) \cdot \frac{\mathcal{F}(\mathbf{z}_t \dot{\oplus} \zeta\mathbf{U}) - \mathcal{F}(\mathbf{z}_t \dot{\oplus} (-\zeta\mathbf{U}))}{2\zeta} \mathbf{U} \quad (14)$$

where ζ is a small perturbation scale (e.g., 10^{-6}) and similarly we define $\mathbf{z}_t \dot{\oplus} \zeta\mathbf{U} \coloneqq [\mathbf{z}_{\mathbf{x},t} + \zeta\mathbf{U}, \mathbf{z}_{\mathbf{h},t}]$ as an abuse of notation. The perturbed representations $[\mathbf{z}_{\mathbf{x},t} \pm \zeta\mathbf{U}, \mathbf{z}_{\mathbf{h},t}]$ do not violate zero center gravity required by t_0, as $\mathbb{E}[\zeta\mathbf{U}] = 0$. Note that, unlike NN guidance, we only add guidance to the positions, i.e., $\mathbf{z}_{\mathbf{x},t}$ and apply no gradient to the atom types, i.e., $\mathbf{z}_{\mathbf{h},t}$. This is because the force definition (Eq. 7) is only physically grounded when the set of atoms stays constant, i.e., no matter/mass is created from or reduced to void. According to Einstein's mass-energy equivalence ($E = mc^2$), any change in atom type would change the mass and create a tremendous potential energy change, which is not within the topics of this work. We present the overall procedure of our non-differentiable oracle guidance in Algorithm 1.

Algorithm 1 Guided diffusion with non-differentiable oracle

input a latent diffusion model ϵ_θ, a VAE decoder $\mathcal{D}$, a composition function $\mathcal{F}$, target property score y, guidance scale s, SPSA perturbation ζ
  $\boldsymbol{z}_T \leftarrow \mathcal{N}(0, \mathbf{I})$
  for all t from T to 1 **do**
    $\mu_{t-1}, \Sigma_{t-1} \leftarrow \frac{1}{1-\beta_t}(\boldsymbol{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t^2}}\epsilon_\theta(\boldsymbol{z}_t, t)), \sigma_t^2\mathbf{I}$
    $\mathbf{U} \leftarrow \mathcal{N}(0, \mathbf{I})$
    $g_{t-1} \propto$
    $-\nabla_{\mathcal{F}(\boldsymbol{z}_t)} \|y - \mathcal{F}(\boldsymbol{z}_t)\|^2 \cdot \left(\frac{\mathcal{F}(\boldsymbol{z}_t \dot{\oplus} \zeta\mathbf{U}) - \mathcal{F}(\boldsymbol{z}_t \dot{\oplus} (-\zeta\mathbf{U}))}{2\zeta}\mathbf{U}\right)$
    (Eq. 14)
    $\boldsymbol{z}_{t-1} \leftarrow \mathcal{N}(\mu_{t-1}, s\Sigma_{t-1}g_{t-1}, \Sigma_{t-1})$ (Eq. 8)
  end for
  $[\boldsymbol{x}, \boldsymbol{h}] \leftarrow \mathcal{D}(\boldsymbol{z}_0)$
output $[\boldsymbol{x}, \boldsymbol{h}]$ //geometry optimized molecules

5. Experiments

5.1. Experiment Setting

Dataset The models in our experiment are trained on the QM9 dataset (Ramakrishnan et al.) and the GEOM dataset (Axelrod & Gómez-Bombarelli). The QM9 dataset (Ramakrishnan et al.) is a catalog with 133,885 small drug-like molecules consisting of up to nine heavy (non-hydrogen) atoms. The Geometric Ensemble Of Molecules (GEOM) dataset (Axelrod & Gómez-Bombarelli) includes 450K molecules with up to 91 heavy atoms (on average, 24.9), where 37 million molecular conformations are generated and reported with their geometries, energies, and statistical weight.

Baseline We apply our method on GeoLDM (Xu et al., 2023), which shows improved performance compared with EDM (Hoogeboom et al., 2022). We compare our method to (unconditional) EDM and GeoLDM, as there are no available ground-truth forces to train their conditional versions.

Evaluation Metric For non-differentiable guidance on forces, We report the Root Mean Square (RMS) and L-1 norm of the forces calculated using xTB as the evaluation metric. We also report results computed using DFT at the B3LYP/6-31G(2df,p) level of theory for some results, a method that is more accurate but exponentially more demanding in computation than GFN2-xTB.

Computation & Implementation We provide computation and implementation details of our method in Appendix A.

5.2. Result

We represent our results in Figure 2, where step N stands for adding guidance in the last N denoising steps. It is observed that over different guidance steps and scales, our non-differentiable oracle guidance can achieve lower forces while generating more valid molecules. Specifically, when the guidance steps begin from the last 400 steps, our method performs the best in terms of lower RMS and L-1 norm net forces (7.22%-9.19% improvement) and higher validity rate (5.0%-10.0% improvement). We provide full numerical results of performance over different guidance steps and scales in Appendix C. To gain further insights, we also report the results on 400 guidance steps computed using DFT in Table 1, a more accurate however expensive calculation. It can be seen that using a less accurate oracle (i.e., xTB) for guidance, we can outperform competitive

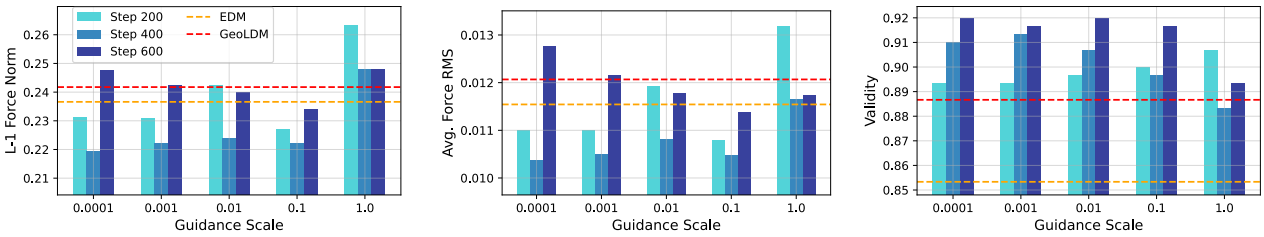

Figure 2. L-1 norm (*left*) and average RMS (*middle*) of forces and validity (*right*) of 300 generated molecules on QM9 with an unconditional GeoLDM guided by xTB, across different guidance steps and scales.

baselines evaluated more precisely (i.e., using DFT), which demonstrates the effectiveness of our method. A similar trend can observed on the GEOM dataset in Table 2, where our best result obtains 6.79% and 6.0% improvements in forces RMS and validity than the best baseline.

Guidance Scale	0.0001	0.001	0.01	0.1	1.0
Force RMS	**0.0051***	0.0052	0.0054	0.0052	0.0061
Validity	96.33%	**96.67%***	96.33%	96.0%	94.0%
EDM	Force RMS 0.0051 / Validity 89.33%				
GeoLDM	Force RMS 0.0061 / Validity 93.67%				

Table 1. Force RMS and validity of 300 generated molecules on QM9 using DFT calculation. $*$ and **bold** denote the overall best and the best within different scales, respectively.

Guidance Scale	0.0001	0.001	0.01	0.1	1.0
Force RMS	**0.0398***	0.0432	0.0451	0.0417	0.0477
Validity	47.0%	48.0%	51.0%	**56.0%***	48.0%
EDM	Force RMS 0.0787 / Validity 45.0%				
GeoLDM	Force RMS 0.0427 / Validity 50.0%				

Table 2. Force RMS and validity of 100 generated molecules on GEOM using xTB calculation. $*$ and **bold** denote the overall best and the best within different scales, respectively.

5.3. Further Discussion

Since xTB runs on CPU and is time costly, we provide further discussion on the effect of the skip-step acceleration method on the xTB guidance in Appendix D.

6. Conclusion & Limitation

In this work, we study conditional generation for 3D diffusion models on molecules, where we steer the generation process of a diffusion model trained unconditionally with an external chemistry oracle, which is usually non-differentiable. By estimating the guidance gradient with an analytic solution, meanwhile, respecting the equivalent and invariant requirements for 3D diffusion models, we can generate molecules with both low net force and high validity. However, since the xTB runs on CPU which is the major bottleneck for inference speed, our method can be time-consuming when the computational resources are insufficient, and future works can explore the possibility of speeding up the guidance steps while reducing computational costs and maintaining good performance.

Impact Statement

To our best knowledge, this work is the first to use a non-differentiable oracle in molecule generation. Our research pioneers a novel approach for incorporating chemistry software into neural networks for a more chemistry-informed and promising generation process and resulting molecules. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Axelrod, S. and Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. 9(1):185. ISSN 2052-4463. doi: 10.1038/s41597-022-01288-4. URL https://www.nature.com/articles/s41597-022-01288-4.

Bannwarth, C., Caldeweyher, E., Ehlert, S., Hansen, A., Pracht, P., Seibert, J., Spicher, S., and Grimme, S. Extended tight-binding quantum chemistry methods. 11(2): e1493. ISSN 1759-0876, 1759-0884. doi: 10.1002/wcms.1493. URL https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1493.

Bannwarth, C., Ehlert, S., and Grimme, S. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 2019.

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

Bao, F., Zhao, M., Hao, Z., Li, P., Li, C., and Zhu, J. Equivariant energy-guided sde for inverse molecular design. In *The eleventh international conference on learning representations*, 2022.

Cao, Y., Romero, J., Olson, J. P., Degroote, M., Johnson, P. D., Kieferová, M., Kivlichan, I. D., Menke, T., Peropadre, B., Sawaya, N. P. D., Sim, S., Veis, L., and Aspuru-Guzik, A. Quantum chemistry in the age of quantum computing. 119(19):10856–10915. ISSN 0009-2665, 1520-6890. doi: 10.1021/acs.chemrev.8b00803. URL https://pubs.acs.org/doi/10.1021/acs.chemrev.8b00803.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

De Cao, N. and Kipf, T. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis, 2021.

Gasteiger, J., Giri, S., Margraf, J. T., and Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In *Machine Learning for Molecules Workshop, NeurIPS*, 2020.

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

Han, X., Shan, C., Shen, Y., Xu, C., Yang, H., Li, X., and Li, D. Training-free multi-objective diffusion model for 3d molecule generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Jin, W., Barzilay, R., and Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Jin, W., Barzilay, R., and Jaakkola, T. Hierarchical graph-to-graph translation for molecules. *arXiv preprint arXiv:1907.11223*, 2019.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.

Luo, Y. and Ji, S. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International Conference on Learning Representations (ICLR)*, 2022.

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., and Arora, S. Fine-tuning language models with just forward passes. *ArXiv*, abs/2305.17333, 2023. URL https://api.semanticscholar.org/CorpusID:258959274.

Maziarz, K., Jackson-Flux, H., Cameron, P., Sirockin, F., Schneider, N., Stiefl, N., Segler, M., and Brockschmidt, M. Learning to extend molecular scaffolds with structural motifs. *arXiv preprint arXiv:2103.03864*, 2021.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Prykhodko, O., Johansson, S. V., Kotsias, P.-C., Arús-Pous, J., Bjerrum, E. J., Engkvist, O., and Chen, H. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, 11(1):1–13, 2019.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. 1(1):140022. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL https://www.nature.com/articles/sdata201422.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Satorras, V. G., Hoogeboom, E., Fuchs, F. B., Posner, I., and Welling, M. E (n) equivariant normalizing flows. *arXiv preprint arXiv:2105.09016*, 2021.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Spall, J. C. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:332–341, 1992. URL https://api.semanticscholar.org/CorpusID:122365276.

Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V., and Frossard, P. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

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

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

A. Computations and Implementations

B. GFN2-xTB

GFN stands for, respectively, **g**eometry optimization, vibrational **f**requencies, and **n**on-covalent interactions. xTB refers to e**x**tended **t**ight **b**inding, and 2 refers to the version. In the GFN2-xTB method, the total energy expression is given by (Bannwarth et al.; 2019)

$$E_{\text{GFN2-xTB}} = E_{\text{rep}} + E_{\text{disp}} + E_{\text{EHT}} + E_{\text{IES+IXC}} + E_{\text{AES}} + E_{\text{AXC}} + G_{\text{Fermi}}$$

, where E_{rep} is the repulsive energy contribution from short-range interactions, E_{disp} is the dispersion energy contribution from long-range interactions, E_{EHT} is the energy contribution from the extended Hückel theory (EHT), $E_{\text{IES+IXC}}$ is the isotropic electrostatic (IES) energy contribution and the isotropic exchange-correlation (IXC) energy contribution, E_{AES} is the anisotropic electrostatic (AES) energy contribution, E_{AXC} is the anisotropic exchange-correlation (AXC) energy contribution, and G_{Fermi} is the entropic contribution of an electronic free energy at finite electronic temperature T_{el} due to Fermi smearing.

Its accuracy and efficiency come strictly from the element-specific and global parameters for all elements up to radon (Z = 86) (Bannwarth et al., 2019), hence the semi-empiricism. The pre-computed tight-binding parameters and empirical corrections are utilized to approximate the electronic structure and calculate energy contributions efficiently.

C. Full Numerical Results

We provide the full numerical results of xTB guided molecule generation in Table 3, where we use guidance steps = [200, 400, 600] and scales = [1e-4, 1e-3, 1e-2, 1e-1, 1.0]. 400 guidance steps and smaller guidance scales give the best results. Note that when guidance step=400 and guidance scale=0.0001, our method outperforms EDM and GeoLDM in the L-1 norm of the forces by 7.23% and 9.22% respectively. The table gives sound evidence that our method generates more stable and valid molecules than the baselines.

Guidance Scale	Guidance Step	L1 Force Norm	Force RMS	Validity
0.0001	200	0.2312	0.0110	89.33%
	400	**0.2195***	**0.0104***	91.0%
	600	0.2477	0.0128	**92.0%***
0.001	200	0.2311	0.0110	89.33%
	400	0.2224	0.0105	91.33%
	600	0.2426	0.0122	91.67%
0.01	200	0.2425	0.0119	89.67%
	400	0.2241	0.0108	90.67%
	600	0.2402	0.0118	**92.0%***
0.1	200	0.2273	0.0108	90.0%
	400	0.2222	0.0105	89.67%
	600	0.2341	0.0114	91.67%
1.0	200	0.2633	0.0132	90.67%
	400	0.2480	0.0117	88.33%
	600	0.2481	0.0117	89.33%
EDM Baseline		0.2366	0.0115	85.33%
GeoLDM Baseline		0.2418	0.0121	88.67%

Table 3. L-1 norm of force, avg. force RMS, and validity from 300 generated molecules sampled from the QM9 dataset using GeoLDM with xTB guidance with different guidance steps and scales using xTB calculation. $*$ and **bold** denote the overall best and the best within different scales, respectively.

D. Acceleration Methods For xTB Guidede Optimization

GPUs can significantly accelerate computation for neural network inference and training, but xTB operates on CPU, which lowers the inference speed. Hence, we propose a *skip-step* acceleration method for xTB guided optimization. Specifically, suppose we set skip-step to be k, then we only calculate gradients from xTB every k steps and use historical gradients for the rest $k-1$ steps. The results are shown in Table 4. For each skip-step schedule, we try the combinations of guidance steps = [200, 4000, 600] and scales = [1e-4, 1e-3]. We can observe that the skip-step method can improve the performance in L-1 norm, RMS of force, and validity, and it is competitive with our previous method without skip-step. Surprisingly, the validity is even higher than our previous method, which we suspect is because our estimation of gradient (Eq. 14) could be noisy and stochastic, so using skip-step method and historical gradients stabilizes the guidance.

Skip Step	Guidance Step	Guidance Scale	L1 Force Norm	Force RMS	Validity
2	200	0.0001	0.2374	0.0114	90.33%
		0.001	0.2306	0.0108	90.0%
	400	0.0001	0.2283	0.0110	**94.0%***
		0.001	**0.2273***	0.0109	93.33%
	600	0.0001	0.2347	0.0113	92.67%
		0.001	0.2426	0.0119	93.67%
3	200	0.0001	0.2274	**0.0107***	88.67%
		0.001	**0.2273***	**0.0107***	88.67%
	400	0.0001	0.2293	0.0109	91.33%
		0.001	0.2297	0.0110	91.33%
	600	0.0001	0.2401	0.0119	92.0%
		0.001	0.2379	0.0116	92.33%
EDM Baseline			0.2366	0.0115	85.33%
GeoLDM Baseline			0.2418	0.0121	88.67%

Table 4. L-1 norm of force, avg. force RMS, and validity from 300 generated molecules sampled from the QM9 dataset using GeoLDM with xTB guidance and various skip-step acceleration schedules using xTB calculation. $*$ and **bold** denote the overall best and the best within different scales, respectively.