Diffusion-based Molecule Generation with Informative Prior Bridges

Chengyue Gong*
University of Texas at Austin
cygong@cs.utexas.edu

Lemeng Wu*
University of Texas at Austin
lmwu@cs.utexas.edu

Xingchao Liu
University of Texas at Austin
xcliu@cs.utexas.edu

Mao Ye
University of Texas at Austin
my21@cs.utexas.edu

Qiang Liu
University of Texas at Austin
lqiang@cs.utexas.edu

Abstract

AI-based molecule generation provides a promising approach to a large area of biomedical sciences and engineering, such as antibody design, hydrolase engineering, or vaccine development. Because the molecules are governed by physical laws, a key challenge is to incorporate prior information into the training procedure to generate high-quality and realistic molecules. We propose a simple and novel approach to steer the training of diffusion-based generative models with physical and statistics prior information. This is achieved by constructing physically informed diffusion bridges, stochastic processes that guarantee to yield a given observation at the fixed terminal time. We develop a Lyapunov function based method to construct and determine bridges, and propose a number of proposals of informative prior bridges for high-quality molecule generation. With comprehensive experiments, we show that our method provides a powerful approach to the 3D generation task, yielding molecule structures with better quality and stability scores.

1 Introduction

As exemplified by the success of AlphafoldV2 [16] in solving protein folding, deep learning techniques have been creating new frontiers on molecular sciences [38]. In particular, the problem of building deep generative models for molecule design has attracted increasing interest with a magnitude of applications in physics, chemistry, and drug discovery [*e.g.*, 1, 2, 19]. Recently, diffusion-based generative model have been applied to molecule generation problems [6, 13] and obtain superior performance. The idea of these methods is to corrupt the data with diffusion noise and learn a neural diffusion model to revert the corruption process to generate meaningful data from noise.

A key challenge in deep generative models for molecule is to efficiently incorporate strong prior information to reflect the physical and problem-dependent statistical properties of the problems at hand. In fact, a recent fruitful line of research [8, 18, 29] have shown promising results by introducing inductive bias into the design of model architectures to reflect physical constraints such as SE(3) equivariance. In this work, we present a different paradigm of prior incorporation tailored to diffusion-based generative models, and leverage it to yield substantial improvement in high-quality and stable molecule generation. Our contributions are summarized as follows.

Prior Guided Learning of Diffusion Models. We introduce a simple and flexible framework for injecting informative problem-dependent prior and physical information when learning diffusion-based generative models. The idea is to elicit and inject prior information regarding how the diffusion

*Equal contribution

NeurIPS 2022 AI for Science Workshop.

process should look like for generating each given data point, and train the neural diffusion model to imitate the prior processes. The prior information is presented in the form of diffusion bridges which are diffusion processes that are guaranteed to generate each data point at the fixed terminal time. We provide a general Lyapunov approach for constructing and determining bridges and leverage it to develop a way to systematically incorporate prior information into bridge processes.

Physics-informed Molecule Generation. We apply our method to molecule generation. We propose a number of energy functions for incorporating physical and statistical prior information. Compared with existing physics-informed molecule generation methods [*e.g.*, 6, 11, 21, 11], our method modifies the training process, rather than imposing constraints on the model architecture. Experiments show that our method achieves current state-of-the-art generation quality and stability on multiple test benchmarks of molecule generation.

2 Related works

Diffuse Bridge Process. Diffusion-based generative models [12, 32, 33, 36] have achieved great successes in various AI generation tasks recently; these methods leverage a time reversion technique and can be viewed as learning variants auto-encoders with diffusion processes as encoders and decoders. Schrodinger bridges [4, 6, 37] have also been proposed for learning diffusion generative models that guarantee to output desirable outputs in a finite time interval, but these methods involve iterative proportional fittings and are computationally costly. Our framework of learning generative models with diffusion bridges is similar to that of [27], which learn diffusion models as a mixture of forward-time diffusion bridges to avoid the time-reversal technique of [35]. But our framework is designed to incorporate physical prior into bridges and develop a systematic approach for constructing a broad class of prior-informed bridges.

3D Molecule Generation. Generating molecule in 3D space has been gaining increasing interest. A line of works [*e.g.* 22, 24, 30, 31, 40, 41, 42] consider conditional conformal generation, which takes the 2D SMILE structure as conditional input and generate the 3D molecule conformations condition on the input. Another series of works [*e.g.*, 10, 13, 20, 29, 39] focus on directly generating the atom position and type for the molecule unconditionally. For these series of works, improvements usually come from architecture design and loss design. For example, G-Schnet [10] auto-regressively generates the atom position and type one by one after another; EN-Flow [29] and EDM [13] adopt E(n) equivariant graph neural network (EGNN) [29] to train flow-based model and diffusion model. These methods aim at generating valid and natural molecules in 3D space and outperform previous approaches by a large margin. Our work provides a very different approach to incorporating the physical information for molecule generation by injecting the prior information into the diffusion process, rather than neural network architectures.

3 Method

We first introduce the definition of diffusion generative models and discuss how to learn these models with prior bridges. After introducing the training algorithm for deep diffusion generative models, we discuss the energy functions that we apply to molecules example.

3.1 Learning Diffusion Generative Models with Prior Bridges

Problem Definition. We aim at learning a generative model given a dataset $\{x^{(k)}\}_{k=1}^n$ drawn from an unknown distribution Π^* on $\mathbb{R}^d$. A diffusion model on time interval $[0,1]$ is

$$\mathbb{P}^\theta: \quad \mathrm{d}Z_t = s_t^\theta(Z_t)\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \quad \forall t \in [0,1], \quad Z_0 \sim \mu_0,$$

where W_t is a standard Brownian motion; $\sigma_t \colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a positive definition covariance coefficient; $s_t^\theta \colon \mathbb{R}^d \to \mathbb{R}^d$ is parameterized as a neural network with parameter θ, and μ_0 is the initialization. Here we use $\mathbb{P}^\theta$ to denote the distribution of the whole trajectory $Z = \{Z_t \colon t \in [0,1]\}$, and $\mathbb{P}_t^\theta$ the marginal distribution of Z_t at time t. We want to learn the parameter θ such that the distribution $\mathbb{P}_1^\theta$ of the terminal state Z_1 equals the data distribution Π^*.

Learning Diffusion Models. There are an infinite number of diffusion processes $\mathbb{P}^\theta$ that yield the same terminal distribution but have different distributions of latent trajectories Z. Hence, it is

important to inject problem-dependent prior information into the learning procedure to obtain a model $\mathbb{P}^\theta$ that simulate the data for the problem at hand fast and accurately. To achieve this, we elicit an *imputation* process $\mathbb{Q}^x$ for each $x \in \mathbb{R}^d$, such that a draw $Z \sim \mathbb{Q}^x$ yields trajectories that 1) are consistent with x in that $Z_1 = x$ deterministically, and 2) reflect important physical and statistical prior information on the problem at hand.

Formally, if $\mathbb{Q}^x(Z_1 = x) = 1$, we call that $\mathbb{Q}^x$ is a bridge process pinned at end point x, or simply an x-bridge. Assume we first generate a data point $x \sim \Pi^*$, and then draw a bridge $Z \sim \mathbb{Q}^x$ pinned at x, then the distribution of Z is a mixture of $\mathbb{Q}^x$ with x drawn from the data distribution: $\mathbb{Q}^{\Pi^*} := \int \mathbb{Q}^x(\cdot)\Pi^*(\mathrm{d}x)$.

A key property of $\mathbb{Q}^{\Pi^*}$ is that its terminal distribution equals the data distribution, i.e., $\mathbb{Q}_1^{\Pi^*} = \Pi^*$. Therefore, we can learn the diffusion model $\mathbb{P}^\theta$ by fitting the trajectories drawn from $\mathbb{Q}^{\Pi^*}$ with the "backward" procedure above. This can be formulated by maximum likelihood or equivalently minimizing the KL divergence:

$$\min_\theta \left\{ \mathcal{L}(\theta) \coloneqq \mathcal{KL}(\mathbb{Q}^{\Pi^*} \mid\mid \mathbb{P}^\theta) \right\}.$$

Furthermore, assume that the bridge $\mathbb{Q}^x$ is a diffusion model of form

$$\mathbb{Q}^x: \quad \mathrm{d}Z_t = b_t(Z_t \mid x)\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \quad Z_0 \sim \mu_0, \tag{1}$$

where $b_t(Z_t \mid x)$ is an x-dependent drift term need to carefully designed to both satisfy the bridge condition and incorporate important prior information (see Section 3.2). Assuming this is done, using Girsanov theorem [25], the loss function $\mathcal{L}(\theta)$ can be reformed into a form ofdenoised score matching loss of [*e.g.*, 33, 35, 34]:

$$\mathcal{L}(\theta) = \mathbb{E}_{Z\sim\mathbb{Q}^{\Pi^*}}\left[\frac{1}{2}\int_0^1 \left\|\sigma(Z_t)^{-1}(s_t^\theta(Z_t) - b_t(Z_t \mid Z_1))\right\|_2^2 \mathrm{d}t\right] + \text{const}, \tag{2}$$

which is a score matching term between s^θ and b. The const term contains the log-likelihood for the initial distribution μ_0, which is a const in our problem. Here θ^* is an global optimum of $\mathcal{L}(\theta)$ if

$$s_t^{\theta^*}(z) = \mathbb{E}_{Z\sim\mathbb{Q}^{\Pi^*}}[b_t(z|Z_1) \mid Z_t = z].$$

This means that the drift term s_t^θ should be matched with the conditional expectation of $b_t(z|x)$ with $x = Z_1$ conditioned on $Z_t = z$.

Remark 3.1. *The SMLD can be viewed as a special case of this framework when we take $\mathbb{Q}^x$ to be a time-scaled Brownian bridge process:*

$$\mathbb{Q}^{x,\mathrm{bb}}: \qquad \mathrm{d}Z_t = \sigma_t^2 \frac{x - Z_t}{\beta_1 - \beta_t}\mathrm{d}t + \sigma_t \mathrm{d}W_t, \quad Z_0 \sim \mathcal{N}(x, \beta_1), \tag{3}$$

where $\sigma_t \in [0, +\infty)$ and $\beta_t = \int_0^t \sigma_s^2 \mathrm{d}s$. This can be seen by the fact that the time-reversed process $\tilde{Z}_t \coloneqq Z_{1-t}$ follows the simple time-scaled Brownian motion $\mathrm{d}\tilde{Z}_t = \sigma_{1-t}\mathrm{d}\tilde{W}_t$ starting from the data point $\tilde{Z}_0 = x$, where $\tilde{W}_t$ is another standard Brownian motion. The Brownian bridge achieves $Z_1 = x$ because the magnitude of the drift force is increasing to infinite when t is close to time 1.

However, the bridge of SMLD above is a relative simple and uninformative process and does not incorporate problem-dependent prior information into the learning procedure. This is also the case of the other standard diffusion-based models [35], such as denoising diffusion probabilistic models (DDPM) which can be shown to use a bridge constructed from an Ornstein–Uhlenbeck process. We refer the readers to [27], which provides a similar forward time bridge framework for learning diffusion models, and it recovers the bridges in SMLD and DDPM as a conditioned stochastic process derived using the h-transform technique [7]. However, the h-transform method is limited to elementary stochastic processes that have an explicit formula of the transition probabilities, and can not incorporate complex physical statistical prior information. Our work strikes to construct and use a broader class of more complex bridge processes that both reflect problem-dependent prior knowledge and satisfy the endpoint condition $\mathbb{Q}^x(Z_1 = x) = 1$. This necessitate systematic techniques for constructing a large family of bridges, as we pursuit in Section 3.2.

3.2 Designing Informative Prior Bridges

The key to realizing the general prior-informed learning framework above is to have a general and user-friendly technique to design $\mathbb{Q}^x$ in (1) to ensure the bridge condition $\mathbb{Q}^x(Z_1 = x) = 1$ while leaving the flexibility of incorporating rich prior information. To achieve this, we first develop a general criterion of bridges based on a *Lyapunov function method* which allows us to identify a very general form of bridge processes; we then propose a particularly simple family of bridges that we use in practice by introducing modification to Brownian bridges.

Definition 3.2 (Lyapunov Functions). *A function $U_t(z)$ is said to be a Lyapunov function for set $A \subset \mathbb{R}^d$ at time $t = 1$ if $U_1(z) \geq 0$ for $\forall z \in \mathbb{R}^d$ and $U_1(z) = 0$ if and only if $z \in A$.*

Intuitively, a diffusion process $\mathbb{Q}$ is a bridge A, i.e., $\mathbb{Q}(Z_1 \in A) = 1$, if it (at least) approximately follows the gradient flow of a Lyapunov function and the magnitude (or step size) or the gradient flow should increase with a proper magnitude in order to ensure that $Z_t \in A$ at the terminal time $t = 1$. Therefore, we identify a general form of bridges to A as follows:

$$\mathbb{Q}^A: \qquad \mathrm{d}Z_t = (-\alpha_t \nabla_z U_t(Z_t) + \nu_t(Z_t))\,\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \qquad t \in [0,1], \quad Z_0 \sim \mu_0, \tag{4}$$

where $\alpha_t > 0$ is the step size of the gradient flow of U and ν is an extra perturbation term. The step size α_t should increase to infinity as $t \to 1$ sufficiently fast to dominate the effect of the diffusion term $\sigma_t \mathrm{d}W_t$ and the perturbation $\nu_t \mathrm{d}t$ term to ensure that U is minimized at time $t = 1$.

Proposition 3.3. *Assume $U_t(z) = U(z,t)$ is a Lyapunov function of a measurable set A at time 1 and $U(\cdot,t) \in C^2(\mathbb{R}^d)$ and $U(z,\cdot) \in C^1([0,1])$. Then, $\mathbb{Q}^A$ in (4) is an bridge to A, i.e., $\mathbb{Q}^A(Z_1 \in A) = 1$, if the following holds:*

1) U follows an (expected) Polyak-Lojasiewicz condition: $\mathbb{E}_{\mathbb{Q}^A}[U_t(Z_t)] - \|\nabla_z U_t(Z_t)\|^2] \leq 0, \forall t$.

2) Let $\beta_t = \mathbb{E}_{\mathbb{Q}^A}[\nabla_z U_t(Z_t)^\top \nu_t(Z_t)]$, and $\gamma_t = \mathbb{E}_{\mathbb{Q}^A}[\partial_t U_t(Z_t) + \frac{1}{2}\mathrm{tr}(\nabla_z^2 U_t(Z_t)\sigma_t^2(Z_t))]$, and $\zeta_t = \exp(\int_0^t \alpha_s \mathrm{d}s)$. Then $\lim_{t\uparrow 1} \zeta_t = +\infty$, and $\lim_{t\uparrow 1} \frac{\zeta_t}{\int_0^t \zeta_s(\beta_s+\gamma_s)\mathrm{d}s} = +\infty$.

Brownian bridge can be viewed as the case when $U_t(z) = \|x - z\|^2/2$ and $\alpha_t = \sigma_t^2/(\beta_1 - \beta_t)$, and $\nu = 0$. Hence simply introducing an extra drift term into bridge bridge yields that a broad family of bridges to x:

$$\mathbb{Q}^{x,\mathrm{bb},f}: \qquad \mathrm{d}Z_t = \left(\sigma_t f_t(Z_t) + \sigma_t^2 \frac{x - Z_t}{\beta_1 - \beta_t}\right)\mathrm{d}t + \sigma_t \mathrm{d}W_t, \quad Z_0 \sim \mu_0. \tag{5}$$

In Appendix A.4 and A.5, we show that $\mathbb{Q}^{x,\mathrm{bb},f}$ is a bridge to x if $\mathbb{E}_{\mathbb{Q}^{x,\mathrm{bb}}}[\|f_t(Z_t)\|^2] < +\infty$ and $\sigma_t > 0, \forall t$, which is very mild condition and is satisfied for most practical functions. The intuition is that the Brownian drift $\sigma_t^2 \frac{x - Z_t}{\beta_1 - \beta_t}$ is singular and grows to infinite as t approaches 1. Hence, introducing an f into the drift would not change of the final bridge condition, unless f is also singular and has a magnitude that dominates the Brownian bridge drift as $t \to 1$.

To make the model $\mathbb{P}^\theta$ compatible with the physical force f, we assume the learnable drift has a form of $s_t^\theta(z) = \alpha f_t(z) + \tilde{s}_t^\theta(z)$ where $\tilde{s}$ is a neural network (typically a GNN) and α can be another learnable parameter or a pre-defined parameter. Please refer to algorithm 3.2 and Figure 1 for descriptions about our practical algorithm.

Algorithm 1 Learning diffusion generative models.

Input: Given a dataset $\{x^{(k)}\}$, $\mathbb{Q}^x$ the bridge in (5), and a problem-dependent prior force f,v and a diffusion model $\mathbb{P}^\theta$.
Training: Estimate θ by minimizing $\mathcal{L}(\theta)$ in (2) with stochastic gradient descent and time discretization.
Sampling: Simulate from $\mathbb{P}^\theta$.

4 Molecule and 3D Generation with Informative Prior Bridges

We apply our method to the molecule generation. Informative physical or statistical priors that reflects the underlying real physical structures can be particularly beneficial for molecule generation as we show in experiments.

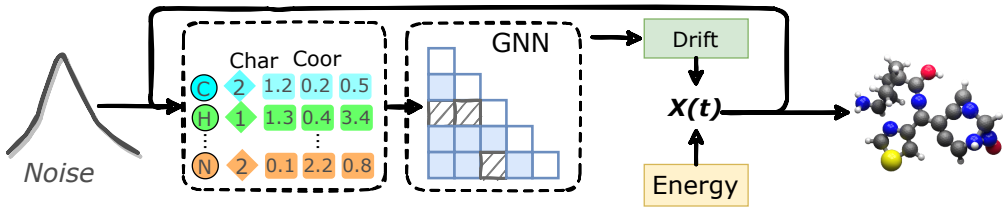

Figure 1: An overview of our training pipeline with molecule generation as an example. Initialized from a given distribution, we pass the data through the network multiple times, and finally get the meaningful output.

In our problem, each data point x is a collection of atoms of different types, more generally marked points, in 3D Euclidean space. In particular, we have $x = [x_i^r, x_i^h]_{i=1}^m$, where $x_i^r \in \mathbb{R}^3$ is the coordinate of the i-th atom, and $x_i^h \in \{e_1, \dots, e_k\}$ where each $e_i = [0 \cdots 1 \cdots 0]$ is the i-th basis vector of $\mathbb{R}^k$, which indicates the type of the i-th atom of k categories. To apply the diffusion generative model, we treat x_i^h as a continuous vector in $\mathbb{R}^r$ and round it to the closest basis vector when we want to output a final result or have computations that depend on atom types (e.g., calculating an energy function as we do in sequel). Specifically, for a continuous $x_i^h \in \mathbb{R}^k$, we denote by $\hat{x}_i^h = \mathbb{I}(x_i^h = \max(x_i^h))$ the discrete type rounded from it by taking the type with the maximum value. To incorporate priors, we design an energy function $E(x)$ and incorporate $f_t(\cdot) = -\nabla E(\cdot)$ into the Brownian bridge (5) to guide the training process. We discuss different choices of E in the following.

4.1 Prior Bridges for Molecule Generation

Previous prior guided molecule or protein 3D structure generation usually depends on pre-defined energy or force [22, 42]. We introduce our two potential energies. One is formulated inspired by previous works in biology, and the other is an k nearest neighbour statistics directly obtained from the data.

AMBER Inspired Physical Energy. AMBER [9] is a family of force fields for molecule simulation. It is designed to provide a computationally efficient tool for modern chemistry-molecular dynamics and free energy calculations. It consists of a number of important forces, including the bond energy, angular energy, torsional energy, the van der Waals energy and the Coulomb energy. Inspired by AMBER, we propose to incorporate the following energy term into the bridge process:

$$E(x) = E_{bond}(x) + E_{angle}(x) + E_{LJ}(x) + E_{Coulomb}(x). \tag{6}$$

• The bond energy is $E_{bond}(x) = \sum_{ij \in bond(x)} (\mathrm{Len}(x_{ij}^r) - \ell(\hat{x}_i^h, \hat{x}_j^h))^2$, where $\mathrm{Len}(x_{ij}^r) = \left\| x_i^r - x_j^r \right\|$, and $bond(x)$ denotes the set of bonds from x, which is set to be the set of atom pairs with a distance smaller than 1.15 times the covalent radius; the $\ell^0(r, c)$ denotes the expected bond length between atom type r and c, which we calculate as side information from the training data.

• The angle energy is $E_{angle}(x) = \sum_{ijk \in angle(x)} (\mathrm{Ang}(x_{ijk}^r) - \omega^0(\hat{x}_{ijk}^h))^2$, where $angle(x)$ denotes the set of angles between two neighbour bonds in $bound(x)$, and $\mathrm{Ang}(x_{ijk}^r)$ denotes the angle formed by vector $x_i^r - x_j^r$ and $x_k^r - x_j^r$, and $\omega^0(\hat{x}_{ijk}^h)$ is the expected angle between atoms of type $\hat{x}_i^h$, $\hat{x}_j^h$, $\hat{x}_k^h$, which we calculate as side information from the training data.

• The Lennard-Jones (LJ) energy is defined by $E_{LJ}(x) = \sum_{i \neq j} e(\left\| x_i^r - x_j^r \right\|)$ and $e(\ell) = (\sigma/\ell)^{12} - 2(\sigma/\ell)^6$. The parameter σ is an approximation for average nucleus distance.

• The nuclei-nuclei repulsion (Coulomb) electromagnetic potential energy is $E_{Coulomb}(x) = \kappa \sum_{ij} q(\hat{x}_i^h) q(\hat{x}_j^h) / \left\| x_i^r - x_j^r \right\|$, where κ is Coulomb constant and $q(r)$ denotes the point charge of atom of type r, which depends on the number of protons.

Statistical Energy. When accurate physic laws are unavailable, molecular geometric statistics, such as bond lengths, bond angles, and torsional angles, etc, can be directly calculated from the data and shed important insights on the system [*e.g.*, 5, 15, 23]. We propose to design a prior energy function in bridges by directly calculate these statistics over the dataset.

Specifically, we assume that the lengths and angles of each type of bond follows a Gaussian distribution that we learn from the dataset, and define the energy function as the negative log-likelihood:

$$E_{stat}(x) = \sum_{ij \in knn(x)} \frac{1}{\hat{\sigma}^2_{\hat{x}^h_{ij}}} \left\| \mathrm{Len}(x^r_{ij}) - \hat{\mu}_{\hat{x}^h_{ij}} \right\|^2 + \sum_{ij,jk \in knn(x)} \frac{1}{\sigma^2_{\hat{x}^h_{ijk}}} \left\| \mathrm{Ang}(x^r_{ijk}) - \mu_{\hat{x}^h_{ijk}} \right\|^2, \quad (7)$$

where $knn(x)$ denotes the K-nearest neighborhood graph constructed based on the distance matrix of x; for each pair of atom types $r, c \in [k]$, $\hat{\mu}_{rc}$ and $\hat{\sigma}^2_{rc}$ denotes empirical mean and variance of length of rc-edges in the dataset; for each triplet $r, c, r' \in [k]$, $\hat{\mu}_{rcr'}$ and $\hat{\sigma}^2_{rcr'}$ is the empirical mean and variance of angle betwen rc and cr' bonds.

Intuitively, depending on the atom type and order of the nearest neighbour, we force the atom distance and angle to mimic the statistics calculated from the data. We thus implicitly capture different kinds of interaction forces. Compared with the AMBER energy, the statistical energy (7) is simpler and more adaptive to the dataset of interest.

5 Experiment

Table 1: Results of our method and several baselines on QM9 and GEOM-DRUG. For QM9, we additionally report the 'Novelty' score evaluated by RDKit [17] to show that our method can generate novel molecules. We evaluate the percentage of valid and unique molecules out of 12000 generated molecules.

	QM9				GEOM-DRUG	
	Atom Sta (%) ↑	Mol Sta (%) ↑	Novelty (%) ↑	Valid + Unique ↑	Atom Sta (%) ↑	Mol Sta (%) ↑
EN-Flow [29]	85.0	4.9	81.4	0.349	75.0	0.0
GDM [13]	97.0	63.2	74.6	-	75.0	0.0
E-GDM [13]	**98.7±0.1**	82.0±0.4	65.7±0.2	0.902	81.3	0.0
Bridge	**98.7±0.1**	81.8±0.2	66.0±0.2	0.902	81.0±0.7	0.0
Bridge + Force (7)	**98.8±0.1**	**84.6±0.3**	68.8±0.2	0.907	**82.4±0.8**	0.0

We verify the advantages of our proposed method (Bridge with Priors) in several different domains. We first compare our method with advanced generators (*e.g.*, diffusion model, normalizing flow, etc.) on molecule generation tasks.

We directly compare the performance and also analyze the difference between our energy prior and other energies we discuss in Section 3.

5.1 Force Guided Molecule Generation

To demonstrate the efficiency and effectiveness of our bridge processes and physical energy, we conduct experiments on molecule and macro-molecule generation experiments. We follow [21] in settings and observe that our proposed prior bridge processes consistently improve the state-of-the-art performance. Diving deeper, we analyze the impact of different energy terms and hyperparameters.

Metrics. Following [13, 29], we use the atom and molecular stability score to measure the model performance. The atom stability is the proportion of atoms that have the right valency while the molecular stability stands for the proportion of generated molecules for which all atoms are stable. For visualization, we use the distance between pairs of atoms and the atom types to predict bond types, which is a common practice. To demonstrate that our force does not only memorize the data in the dataset, we further calculate and report the RDKit-based [17] novelty score. we extracted 10,000 samples to calculate the above metrics.

Dataset Settings QM9 [28] molecular properties and atom coordinates for 130k small molecules with up to 9 heavy atoms with 5 different types of atoms. This data set contains small amino acids, such as GLY, ALA, as well as nucleobases cytosine, uracil, and thymine. We follow the common practice in [13] to split the train, validation, and test partitions, with 100K, 18K, and 13K samples. GEOM-DRUG [3] is a dataset that contains drug-like molecules. It features 37 million molecular conformations annotated by energy and statistical weight for over 450,000 molecules. Each molecule contains 44 atoms on average, with 5 different types of atoms. Following [13, 29], we retain the 30 lowest energy conformations for each molecule.

Training Configurations. On QM9, we train the EGNNs with 256 hidden features and 9 layers for 1100 epochs, a batch size 64, and a constant learning rate 10^{-4}, which is the default training configuration. We use the polynomial noise schedule used in [13] which linearly decay from $10^{-2}/T$

Table 2: We compare w. and w/o force results with different discretization time steps.

	Time Step					
	50		100		500	
	Atom Stable (%)	Mol Stable (%)	Atom Stable (%)	Mol Stable (%)	Atom Stable (%)	Mol Stable (%)
EGM	97.0±0.1	66.4±0.2	97.3±0.1	69.8±0.2	**98.5±0.1**	**81.2±0.1**
Bridge + Force (7)	**97.3±0.1**	**69.2±0.2**	**97.9±0.1**	**72.3±0.2**	**98.7±0.1**	**83.7±0.1**

to 0. We linearly decay α from $10^{-3}/T$ to 0 *w.r.t.* time step. We set $k = 5$ (7) by default. On GEOM-DRUG, we train the EGNNs with 256 hidden features and 8 layers with batch size 64, a constant learning rate 10^{-4}, and 10 epochs. It takes approximately 10 days to train the model on these two datasets on one `Tesla V100-SXM2-32GB` GPU. We provide E(3) Equivariant Diffusion Model (EDM) [13] and E(3) Equivariant Normalizing Flow (EN-Flow) [29] as our baselines. Both two are trained with the same configurations as ours.

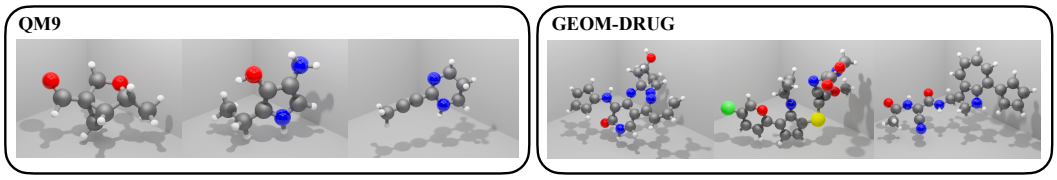

Figure 2: Examples of molecules generated by our method on QM9 and GEOM-DRUG.

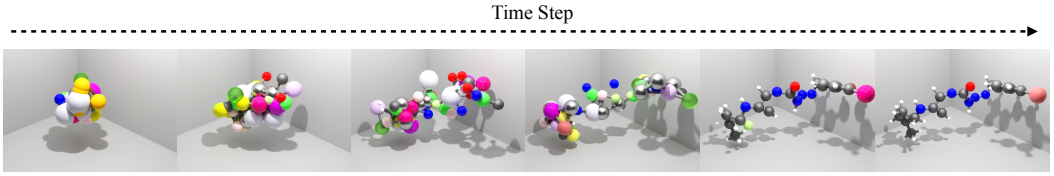

Figure 3: An example of generation trajectory following $\mathbb{P}^\theta$ of our method, trained on GEOM-DRUG.

Results: Higher Quality and Novelty. We summarize our experimental results in Table 1. We observe that **(1)** our method generates molecules with better qualities than the others. On QM9, we notice that we improve the molecule stability score by a large margin (from 82.0 to 84.6) and slightly improve the atom stability score (from 98.7 to 98.8). It indicates that with the informed prior bridge helps improves the quality of the generated molecules. **(2)** Our method achieves a better novelty score. Compared to E-GDM, we improve the novelty score from 65.7 to 68.8. This implies that our introduced energy does not hurt the novelty when the statistics are estimated over the training dataset. Notice that although the GDM and EN-Flow achieve a better novelty score, the sample quality is much worse. The reason is that, due to the metric definition, low-quality out-of-distribution samples lead to high novelty scores. **(3)** On the GEOM-DRUG dataset, the atom stability is improved from 81.3 to 82.4, which shows that our method can work for macro-molecules. **(4)** We visualize and qualitatively evaluate our generate molecules. Figure 3 displays the trajectory on GEOM-DRUG and Figure 2 shows the samples on two datasets. **(5)** Bridge processes and E-GDM obtain comparable results on our tested benchmarks. **(6)** The computational load added by introducing prior bridges is small. Compared to EGM, we only introduce 8% additional cost in training and 3% for inference.

Result: Better With Fewer Time Steps. We display the performance of our method with fewer time steps in Table 2. We observe that **(1)** with fewer time steps, the baseline EGM method gets worse results than 1000 steps in Table 1. **(2)** with 500 steps, our method still keeps a consistently good performance. **(3)** with even fewer 50 or 100 steps, our method yields a worse result than 1000 steps in Table 1, but still outperforms the baseline method by a large margin.

Table 3: We compare EGM models trained with different force mentioned in Section 3.

Method	Atom Stable (%)	Mol Stable (%)	Method	Atom Stable (%)	Mol Stable (%)
Force (7), $k = 7$	**98.8±0.1**	**84.5±0.2**	Force (6)	98.7±0.1	83.1±0.2
Force (7), $k = 5$	**98.8±0.1**	**84.6±0.3**	Force (6) w/o. bond	98.7±0.1	82.5±0.1
Force (7), $k = 3$	**98.8±0.1**	83.9±0.3	Force (6) w/o. angle	98.7±0.1	82.4±0.2
Force (7), $k = 1$	**98.8±0.1**	82.7±0.3	Force (6) w/o. Long-range	98.7±0.1	82.7±0.2

Ablation: Impacts of Different Energies. We apply several energies we discuss in Section 3, and compare them on the QM9 dataset. **(1)** We notice that our energy (7) gets better performance with larger k when $k \leq 5$. $k = 7$ achieves comparable performance as $k = 5$. Larger k also requires more computation time, which yields a trade-off between performance and efficiency. **(2)** For (6), once removing a typical term, the performance drops. **(3)** In all the cases, applying additional forces outperforms the bridge processes baseline w/o. force.

6 Conclusion and Limitations

We propose a framework to inject informative priors into learning neural parameterized diffusion models, with applications to both molecules and 3D point cloud generation. Empirically, we demonstrate that our method has the advantages such as better generation quality, less sampling time and easy-to-calculate potential energies. For future works, we plan to 1) study the relation between different types of forces for different domain of molecules, 2) study how to generate valid proteins in which the number of atoms is very large, and 3) apply our method to more realistic applications such as antibody design or hydrolase engineering.

In both energy functions in (7) and (6), we do not add torsional angle related energy [14] mainly because it is hard to verify whether four atoms are bonded together during the stochastic process. We plan to study how to include this for better performance in future works.

Another weakness of deep diffusion bridge processes are their computation time. Similar to previous diffusion models [21], it takes a long time to train a model. We attempted to speed the training up by using a large batch size (*e.g.*, 512, 1024) but found a performance drop. An important future direction is to study methods to distribute and accelerate the training.

References

[1] Miguel Alcalde, Manuel Ferrer, Francisco J Plou, and Antonio Ballesteros. Environmental biocatalysis: from remediation with enzymes to novel green processes. *TRENDS in Biotechnology*, 24(6):281–287, 2006.

[2] Namrata Anand, Raphael Eguchi, Irimpan I Mathews, Carla P Perez, Alexander Derry, Russ B Altman, and Po-Ssu Huang. Protein sequence design with a learned potential. *Nature communications*, 13(1):1–11, 2022.

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

[4] Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood training of schr\" odinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*, 2021.

[5] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[6] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Joseph L Doob and JI Doob. *Classical potential theory and its probabilistic counterpart*, volume 549. Springer, 1984.

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

[9] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012, 2003.

[10] Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7566–7578. Curran Associates, Inc., 2019.

[11] Dwaraknath Gnaneshwar, Bharath Ramsundar, Dhairya Gandhi, Rachel Kurchin, and Venkatasubramanian Viswanathan. Score-based generative models for molecule generation. *arXiv preprint arXiv:2203.04698*, 2022.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[13] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. *arXiv preprint arXiv:2203.17003*, 2022.

[14] Bowen Jing, Gabriele Corso, Regina Barzilay, and Tommi S Jaakkola. Torsional diffusion for molecular conformer generation. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

[15] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[16] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[17] Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

[18] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

[19] Hongyuan Lu, Daniel J Diaz, Natalie J Czarnecki, Congzhi Zhu, Wantae Kim, Raghav Shroff, Daniel J Acosta, Bradley R Alexander, Hannah O Cole, Yan Zhang, et al. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907):662–667, 2022.

[20] Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d molecule generative model for structure-based drug design. *arXiv preprint arXiv:2203.10446*, 2022.

[21] Shitong Luo, Jiahan Li, Jiaqi Guan, Yufeng Su, Chaoran Cheng, Jian Peng, and Jianzhu Ma. Equivariant point cloud analysis via learning orientations for message passing. *arXiv preprint arXiv:2203.14486*, 2022.

[22] Shitong Luo, Chence Shi, Minkai Xu, and Jian Tang. Predicting molecular conformation via dynamic graph score matching. *Advances in Neural Information Processing Systems*, 34, 2021.

[23] M Riad Manaa, Laurence E Fried, Carl F Melius, Marcus Elstner, and Th Frauenheim. Decomposition of hmx at extreme conditions: A molecular dynamics simulation. *The Journal of Physical Chemistry A*, 106(39):9024–9029, 2002.

[24] Elman Mansimov, Omar Mahmood, Seokho Kang, and Kyunghyun Cho. Molecular geometry prediction using a deep generative graph neural network. *Scientific reports*, 9(1):1–13, 2019.

[25] Xuerong Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

[26] B. Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 6 edition, 2013.

[27] Stefano Peluchetti. Non-denoising forward-time diffusions, 2022.

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

[29] Victor Garcia Satorras, Emiel Hoogeboom, Fabian B Fuchs, Ingmar Posner, and Max Welling. E (n) equivariant normalizing flows. *arXiv preprint arXiv:2105.09016*, 2021.

[30] Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, 2021.

[31] Gregor NC Simm and José Miguel Hernández-Lobato. A generative model for molecular distance geometry. *arXiv preprint arXiv:1909.11459*, 2019.

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

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[34] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34, 2021.

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[36] Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

[37] Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via schrödinger bridge. In *International Conference on Machine Learning*, pages 10794–10804. PMLR, 2021.

[38] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, 13(1):e1005324, 2017.

[39] Fang Wu, Qiang Zhang, Xurui Jin, Yinghui Jiang, and Stan Z Li. A score-based geometric model for molecular dynamics simulations. *arXiv preprint arXiv:2204.08672*, 2022.

[40] Minkai Xu, Shitong Luo, Yoshua Bengio, Jian Peng, and Jian Tang. Learning neural generative dynamics for molecular conformation generation. *arXiv preprint arXiv:2102.10240*, 2021.

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

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

A Proofs

Proof of Proposition 3.3. It is a direct result of Theorem A.1. □

Theorem A.1. *Assume*

$$\mathrm{d}Z_t = \eta(Z_t, t)\mathrm{d}t + \sigma(Z_t, t)\mathrm{d}W_t, \qquad t \in [0, 1].$$

We have $Z_1 \in A$ with probability one if there exists a function $U\colon \mathbb{R}^d \times [0,1] \to \mathbb{R}$ such that

1) $U(\cdot, t) \in C^2(\mathbb{R}^d)$ and $U(z, \cdot) \in C^1([0, 1])$;

2) $U(z, 1) \geq 0$, $z \in \mathbb{R}^d$, and $U(z, 1) = 0$ implies that $z \in A$, where A is a measurable set in $\mathbb{R}^d$;

3) There exists a sequence $\{\alpha_t, \beta_t, \gamma_t\colon t \in [0, 1]\}$, such that for $t \in [0, 1]$,

$$\mathbb{E}[\nabla_z U(Z_t, t)^\top \eta(Z_t, t)] \leq -\alpha_t \mathbb{E}[U(Z_t, t)] + \beta_t,$$
$$\mathbb{E}[\partial_t U(Z_t, t) + \frac{1}{2}\mathrm{tr}(\nabla_z^2 U(Z_t, t)\sigma^2(Z_t, t))] \leq \gamma_t;$$

4) Define $\zeta_t = \exp(\int_0^t \alpha_s \mathrm{d}s)$. We assume

$$\lim_{t\uparrow T} \zeta_t = +\infty, \quad \lim_{t\uparrow T} \frac{\zeta_t}{\int_0^t \zeta_s(\beta_s + \gamma_s)\mathrm{d}s} = +\infty. \tag{8}$$

Proof. Following $\mathrm{d}Z_t = \eta(Z_t, t)\mathrm{d}t + \sigma(Z_t, t)\mathrm{d}W_t$, we have by Ito's Lemma,

$$\mathrm{d}U(Z_t, t) = \nabla U(Z_t, t)^\top(\eta(Z_t, t)\mathrm{d}t + \sigma(Z_t, t)\mathrm{d}W_t) + \partial_t U(Z_t, t)\mathrm{d}t + \frac{1}{2}\mathrm{tr}(\nabla^2 U(Z_t, t)\sigma^2(Z_t, t))\mathrm{d}t,$$

for $t \in [0, T]$. Taking expectation on both sides,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}(U(Z_t)) = \mathbb{E}[\nabla_z U(Z_t, t)^\top \eta(Z_t, t)] + \mathbb{E}\left[\partial_t U(Z_t, t) + \frac{1}{2}\mathrm{tr}(\nabla^2 U(Z_t, t)\sigma^2(Z_t, t))\right].$$

Let $u_t = \mathbb{E}[U(Z_t, t)]$. By the assumption above, we get

$$\dot{u}_t \leq -\alpha_t u_t + \beta_t + \gamma_t.$$

Following Grönwall's inequality (see Lemma A.2 below), we have $\mathbb{E}[U(Z_1, 1)] = u_1 = \lim_{t\uparrow 1} u_t \leq 0$ if (8) holds. Because $U(z, 1) \geq 0$, this suggests that $U(Z_1, 1) = 0$ and hence $Z_1 \in A$ almost surely. □

Lemma A.2. *Let $u_t \in \mathbb{R}$ and $\alpha_t, \beta_t \geq 0$, and $\frac{\mathrm{d}}{\mathrm{d}t}u_t \leq -\alpha_t u_t + \beta_t$, $t \in [0, T]$ for $T > 0$. We have*

$$u_t \leq \frac{1}{\zeta_t}(\zeta_0 u_0 + \int_0^t \zeta_s \beta_s \mathrm{d}s), \qquad \textit{where} \qquad \zeta_t = \exp(\int_0^t \alpha_s \mathrm{d}s).$$

Therefore, we have $\lim_{t\uparrow T} u_t \leq 0$ if

$$\lim_{t\uparrow T} \zeta_t = +\infty, \quad \lim_{t\uparrow T} \frac{\zeta_t}{\int_0^t \zeta_s \beta_s \mathrm{d}s} = +\infty.$$

Proof. Let $v_t = \zeta_t u_t$, where $\zeta_t = \exp(\int_0^t \alpha_s \mathrm{d}s)$ so $\dot{\zeta}_t = \zeta_t \alpha_t$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}v_t = \dot{\zeta}_t u_t + \zeta_t \dot{u}_t \leq (\dot{\zeta}_t - \zeta_t \alpha_t)u_t + \zeta_t \beta_t = \zeta_t \beta_t.$$

So

$$v_t \leq v_0 + \beta \int_0^t \gamma_s \mathrm{d}s,$$

and hence

$$u_t \leq \frac{1}{\zeta_t}(\zeta_0 u_0 + \int_0^t \zeta_s \beta_s \mathrm{d}s).$$

To make $\lim_{t\uparrow T} u_t \leq 0$, we want

$$\lim_{t\uparrow T} \zeta_t = +\infty, \quad \lim_{t\uparrow T} \frac{\zeta_t}{\int_0^t \zeta_s \beta_s \mathrm{d}s} = +\infty.$$

□

Corollary A.3. *Let* $\mathrm{d}Z_t = \frac{x-Z_t}{1-t} + \varsigma_t \mathrm{d}W_t$ *with law* $\mathbb{Q}$*. This uses the drift term of Brownian bridge, but have a time-varying diffusion coefficient* $\varsigma_t \geq 0$*. Assume* $\sup_{t\in[0,T]} \varsigma_t < \infty$*. Then* $\mathbb{Q}(Z_1 = z) = 1$*.*

Proof. We verify the conditions in Theorem A.1. Define $U(z,t) = \|x-z\|^2/2$, and $\eta(z,t) = \frac{x-Z_t}{1-t}$. We have $\eta(z,t)^\top \nabla U(z,t) = -U(z,t)/(T-t)$. So $\alpha_t = 1/(T-t)$.

Also, $\partial_t U(z,t) + \frac{1}{2}\mathrm{tr}(\varsigma_t^2 \nabla_z^2 U(z,t)) = \frac{1}{2}\mathrm{diag}(\varsigma_t^2 I_{d\times d}) = \frac{d}{2}\varsigma_t^2 \coloneqq \beta_t \leq C < \infty$.

Then $\zeta_t = \exp(\int_0^t \alpha_s \mathrm{d}s) = \frac{1}{1-t} \to +\infty$ as $t \uparrow T$.

Also, $\int_0^t \zeta_s \beta_s \mathrm{d}s \leq C \int_0^t \zeta_s \mathrm{d}s = CT(\log(T) - \log(T-t))$. So

$$\lim_{t\uparrow T} \frac{\zeta_t}{\int_0^t \zeta_s \beta_s \mathrm{d}s} \geq \lim_{t\uparrow T} \frac{\frac{1}{1-t}}{CT(\log(T) - \log(T-t))} = +\infty.$$

□

Using Girsanov theorem, we show that introducing arbitrary non-singular changes (as defined below) on the drift and initialization of a process does not change its bridge conditions.

Proposition A.4. *Consider the following processes*

$$\begin{aligned} \mathbb{Q}\colon &\quad Z_t = b_t(Z_t)\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \quad Z_0 \sim \mu_0 \\ \tilde{\mathbb{Q}}\colon &\quad Z_t = (b_t(Z_t) + \sigma_t(Z_t) f_t(Z_t))\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \quad Z_0 \sim \tilde{\mu}_0. \end{aligned}$$

Assume we have $\mathcal{KL}(\mu_0 \,||\, \tilde{\mu}_0) < +\infty$ *and* $\mathbb{E}_{\mathbb{Q}}[\int_0^T \|f_t(Z_t)\|^2] < \infty$*. Then for any event* A*, we have* $\mathbb{Q}(Z \in A) = 1$ *if and only if* $\tilde{\mathbb{Q}}(Z \in A) = 1$*.*

Proof. Using Girsnaov theorem [26], we have

$$\mathcal{KL}(\mathbb{Q} \,||\, \tilde{\mathbb{Q}}) = \mathcal{KL}(\mu_0 \,||\, \tilde{\mu}_0) + \frac{1}{2}\mathbb{E}_{\mathbb{Q}}\left[\int_0^1 \|f_t(Z_t)\|_2^2 \,\mathrm{d}t\right].$$

Hence, we have $\mathcal{KL}(\mathbb{Q} \,||\, \tilde{\mathbb{Q}}) < +\infty$. This implies that $\mathbb{Q}$ and $\tilde{\mathbb{Q}}$ has the same support. Hence $\mathbb{Q}(Z \in A) = 1$ iff $\tilde{\mathbb{Q}}(Z \in A) = 1$ for any measurable set A. □

This gives an immediate proof of the following result that we use in the paper.

Corollary A.5. *Consider the following two processes:*

$$\begin{aligned} \mathbb{Q}^{x,\mathrm{bb}}\colon &\quad \mathrm{d}Z_t = \left(\sigma_t^2 \frac{x - Z_t}{\beta_1 - \beta_t}\right)\mathrm{d}t + \sigma_t \mathrm{d}W_t, \quad Z_0 \sim \mu_0, \\ \mathbb{Q}^{x,\mathrm{bb},f}\colon &\quad \mathrm{d}Z_t = \left(\sigma_t f_t(Z_t) + \sigma_t^2 \frac{x - Z_t}{\beta_1 - \beta_t}\right)\mathrm{d}t + \sigma_t \mathrm{d}W_t, \quad Z_0 \sim \mu_0. \end{aligned}$$

Assume $\mathbb{E}_{\mathbb{Q}^{x,\mathrm{bb},f}}[\|f_t(Z_t)\|^2] < +\infty$ *and* $\sigma_t > 0$ *for* $t \in [0, +\infty)$*. Then* $\mathbb{Q}^{x,\mathrm{bb},f}$ *is a bridge to* x*.*

B Model Details

B.1 Model Architecture for Molecule Generation.

Following EGM [13], we apply an E(3) equivariant GNN network (EGNN) as our basic model architecture. EGNNs are a type of graph neural networks that satisfies the equivariance constraint,

$$\mathbf{R}x' + \mathbf{t}, h' = f(\mathbf{R}x + \mathbf{t}, h) \quad \textit{when} \quad x', h' = f(x, h), \tag{9}$$

where x and h represent the 3D coordinates and additional features, orthogonal $\mathbf{R}$ stands for the random rotation and $\mathbf{t} \in \mathbb{R}^3$ is a random transformation. One EGNN is usually made up of multiple stacked equivariant graph convolutional layers (EGCL), and every EGCL satisfies the

equivariance constraint. Denote N the number of nodes, x^l and h^l the coordinates and features for layer $l \in \{0, \cdots, L\}$, we have

$$\begin{aligned} m_{ij} &= \phi_e(h_i^l, h_j^l, d_{ij}), \\ h_i^{l+1} &= \phi_h(h_i^l, \{m_{ij}\}_{j=1}^N), \\ x_i^{l+1} &= x_i^l + \sum_{j \neq i} \frac{x_i^l - x_j^l}{d+1} \phi_x(h_i^l, h_j^l, d_{ij}), \end{aligned} \tag{10}$$

where $h^0 = h, x^0 = x$, $d_{ij} = \|x_i^l - x_j^l\|_2$, $d_{ij} + 1$ is introduced to improve training stability, and ϕ_e, ϕ_h, ϕ_x represents fully connected neural network with learnable parameters. We refer the readers to the previous paper [29] for more details.

Scaling Features Following [13], we re-scale the data with additional scaling factors. The atom type one-hot vector and atom charge value $\times .25$ and $\times 0.1$, respectively. It significantly improves performance over non-scaled inputs, e.g. 47% relative improvements on molecule stability.