DIFFMD: A Geometric Diffusion Model for Molecular Dynamics Simulations

Fang Wu[1,2], **Stan Z. Li**[1*]

[1] AI Research and Innovation Laboratory, School of Engineering, Westlake University
[2] Institute of AI Industry Research, Tsinghua University
fw2359@columbia.edu, stan.zq.li@westlake.edu.cn

Abstract

Molecular dynamics (MD) has long been the *de facto* choice for simulating complex atomistic systems from first principles. Recently deep learning models become a popular way to accelerate MD. Notwithstanding, existing models depend on intermediate variables such as the potential energy or force fields to update atomic positions, which requires additional computations to perform back-propagation. To waive this requirement, we propose a novel model called DIFFMD by directly estimating the gradient of the log density of molecular conformations. DIFFMD relies on a score-based denoising diffusion generative model that perturbs the molecular structure with a conditional noise depending on atomic accelerations and treats conformations at previous timeframes as the prior distribution for sampling. Another challenge of modeling such a conformation generation process is that a molecule is kinetic instead of static, which no prior works have strictly studied. To solve this challenge, we propose an equivariant geometric Transformer as the score function in the diffusion process to calculate corresponding gradients. It incorporates the directions and velocities of atomic motions via 3D spherical Fourier-Bessel representations. With multiple architectural improvements, we outperform state-of-the-art baselines on MD17 and isomers of $C_7O_2H_{10}$ datasets. This work contributes to accelerating material and drug discovery.

Introductions

Molecular dynamics (MD), an *in silico* tool that simulates complex atomic systems based on first principles, has exerted dramatic impacts in scientific research. Instead of yielding an average structure by experimental approaches including X-ray crystallography and cryo-EM, MD simulations can capture the sequential behavior of molecules in full atomic details at the very fine temporal resolution, and thus allow researchers to quantify how much various regions of the molecule move at equilibrium and what types of structural fluctuations they undergo. In the areas of molecular biology and drug discovery, the most basic and intuitive application of MD is to assess the mobility or flexibility of various regions of a molecule. MD substantially accelerates the studies to observe the biomolecular processes in action, particularly important functional processes such as ligand binding (Shan et al. 2011), ligand- or voltage-induced conformational change (Dror et al. 2011), protein folding (Lindorff-Larsen et al. 2011), or membrane transport (Suomivuori et al. 2017).

Nevertheless, the computational cost of MD generally scales cubically with respect to the number of electronic degrees of freedom. Besides, important biomolecular processes like conformational change often take place on timescales longer than those accessible by classical all-atom MD simulations. Although a wide variety of enhanced sampling techniques have been proposed to capture longer-timescale events (Schwantes, McGibbon, and Pande 2014), none of them is a panacea for timescale limitations and might additionally cause decreased accuracy. Thus, it is an urgent demand to fundamentally boost the efficiency of MD while keeping accuracy.

Recently, deep learning-based MD (DLMD) models provide a new paradigm to meet the pressing demand. The accuracy of those models stems from not only the distinctive ability of neural networks to approximate high-dimensional functions but the proper treatment of physical requirements like symmetry constraints and the concurrent learning scheme that generates a compact training dataset (Jia et al. 2020). Despite their success, current DLMD models primarily suffer from the following three issues. First, most DLMD models still rely on intermediate variables (e.g., the potential energy) and multiple stages to generate subsequent biomolecular conformations. This substantially raises the computational expenditure and hinders the inference efficiency, since the inverse Hessian scales as cubically with the number of atom coordinates (Cranmer et al. 2020). Second, existing DLMD models regard the DL module as a black-box to predict atomic attributes and never inosculate the neural architecture with the theory of thermodynamics. Last but not least, the majority of prevailing geometric methods (Gilmer et al. 2017; Schütt et al. 2018; Klicpera, Groß, and Günnemann 2020) are designed for immobile molecules and not suitable for dynamic systems where the directions and velocities of atomic motions count.

This paper proposes DIFFMD that aims to address the above-mentioned issues. First, DIFFMD is a one-stage procedure and forecasts the simulation trajectories without any dependency on the potential energy or forces. For the sec-

*The corresponding author.

ond issue, inspired by the consistency of diffusion processes in nonequilibrium thermodynamics and probabilistic generative models (Sohl-Dickstein et al. 2015; Song and Ermon 2019), DIFFMD adopts a score-based denoising diffusion generative model (Song et al. 2020b) with the exploration of various stochastic differential equations (SDEs). It sequentially corrupts training data by slowly increasing noise and then learns to reverse this corruption. This generative process highly accords with the enhanced sampling mechanism in MD (Miao, Feher, and McCammon 2015), where a boost potential is added conditionally to smooth biomolecular potential energy surface and decrease energy barriers. Besides, to make geometric models aware of atom mobility, we propose an equivariant geometric Transformer (EGT) as the score function for our DIFFMD. It refines the self-attention mechanism (Vaswani et al. 2017) with 3D spherical Fourier-Bessel representations to incorporate both the intersection and dihedral angles between each pair of atoms and their associated velocities.

We conduct comprehensive experiments on multiple standard MD simulation datasets including MD17 and $C_7O_2H_{10}$ isomers. Numerical results demonstrate that DIFFMD constantly outperforms state-of-the-art DLMD models by a large margin. The significantly superior performance illustrates the high capability of our DIFFMD to accurately produce MD trajectories for microscopic systems.

Preliminaries

Background

We consider an MD trajectory of a molecule with T timeframes. $\mathcal{M}^{(t)} = \left(\boldsymbol{x}^{(t)}, \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t)}\right)$ denotes the conformation $\boldsymbol{x}^{(t)}$ at time $t \in [T]$ and is assumed to have N atoms. There $\boldsymbol{x}^{(t)} \in \mathbb{R}^{N\times 3}$ and $\boldsymbol{h}^{(t)} \in \mathbb{R}^{N\times \psi_h}$ denote the 3D coordinates and ψ_h-dimension roto-translational invariant features (e.g. atom types) associated with each atom, respectively. $\boldsymbol{v}^{(t)} \in \mathbb{R}^{N\times 3}$ corresponds to the atomic velocities. We denote a vector norm by $x = \|\boldsymbol{x}\|_2$, its direction by $\hat{\boldsymbol{x}} = \boldsymbol{x}/x$, and the relative position by $\boldsymbol{x}_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j$.

Molecular Dynamics

MD with classical potentials. The fundamental idea behind MD simulations is to study the time-dependent behavior of a microscopic system. It generates the atomic trajectories for a specific interatomic potential with certain initial conditions and boundary conditions. This is obtained by solving the first-order differential equation of the Newton's second law:

$$\boldsymbol{F}_i^{(t)} = m_i \boldsymbol{a}_i^{(t)} = -\frac{\partial U\left(\boldsymbol{x}^{(t)}\right)}{\partial \boldsymbol{x}_i^{(t)}}, \tag{1}$$

where $\boldsymbol{F}_i^{(t)}$ is the net force acting on the i-th atom of the system at a given point in the t-th timeframe, $\boldsymbol{a}_i^{(t)}$ is the corresponding acceleration, and m_i is the mass. $U(\boldsymbol{x})$ is the potential energy function. The classic force field (FF) defines the potential energy function in Appendix. Then numerical methods are utilized to advance the trajectory over small time increments Δt with the assistance of some integrator (see more introductions to MD in Appendix).

Enhanced sampling in MD. Enhanced sampling methods have been developed to accelerate MD and retrieve useful thermodynamic and kinetic data (Rocchia, Masetti, and Cavalli 2012). These methods exploit the fact that the free energy is a state function; thus, differences in free energy are independent of the path between states (De Vivo et al. 2016). Several techniques such as free-energy perturbation, umbrella sampling, tempering, and metadynamics are invented to reduce the energy barrier and smooth the potential energy surface (Luo et al. 2020; Liao 2020).

Score-based Generative Model

Score-based generative models (Song et al. 2020b) refer to the score matching with Langevin dynamics (Song and Ermon 2019) and the denoising diffusion probabilistic modeling (Sohl-Dickstein et al. 2015) .They have shown effectiveness in the generation of images (Ho, Jain, and Abbeel 2020) and molecular conformations (Shi et al. 2021).

Diffusion process. Assume a diffusion process $\{\boldsymbol{x}(s)\}_{s=0}^{S}$ indexed by a continuous time variable $s \in [0, S]$, such that $\boldsymbol{x}(0) \sim p_0$, for which we have a dataset of i.i.d. samples, and $\boldsymbol{x}(S) \sim p_S$, for which we have a tractable form to generate samples efficiently. Let $p_s(\boldsymbol{x})$ be the probability density of $\boldsymbol{x}(s)$, and $p(\boldsymbol{x}(s_1) \mid \boldsymbol{x}(s_0))$ be the transition kernel from $\boldsymbol{x}(s_0)$ to $\boldsymbol{x}(s_1)$, where $0 \le s_0 < s_1 \le T$. Then the diffusion process is modeled as the solution to an Itô SDE (Song et al. 2020b):

$$\mathrm{d}\boldsymbol{x} = f(\boldsymbol{x}, s)\mathrm{d}s + g(s)\mathrm{d}\boldsymbol{w}, \tag{2}$$

where $\boldsymbol{w}$ is a standard Wiener process, $f(\cdot, s) : \mathbb{R}^d \to \mathbb{R}^d$ is a vector-valued function called the drift coefficient of $\boldsymbol{x}(s)$, and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is a scalar function known as the diffusion coefficient of $\boldsymbol{x}(s)$.

Reverse process. By starting from samples of $\boldsymbol{x}(S) \sim p_S$ and reversing the diffusion process, we can obtain samples $\boldsymbol{x}(0) \sim p_0$. The reverse-time SDE can be acquired based on the result from Anderson (1982) that the reverse of a diffusion process is also a diffusion process as:

$$\mathrm{d}\boldsymbol{x} = \left[f(\boldsymbol{x}, s) - g(s)^2 \nabla_{\boldsymbol{x}} \log p_s(\boldsymbol{x})\right] \mathrm{d}s + g(s)\mathrm{d}\overline{\boldsymbol{w}}, \tag{3}$$

where $\overline{\boldsymbol{w}}$ is a standard Wiener process when time flows backwards from S to 0, and $\mathrm{d}s$ is an infinitesimal negative timeframe. The score of a distribution can be estimated by training a score-based model on samples with score matching (Song and Ermon 2019). To estimate $\nabla_{\boldsymbol{x}} \log p_s(\boldsymbol{x})$, one can train a time-dependent score-based model $\boldsymbol{s}_{\boldsymbol{\vartheta}}(\boldsymbol{x}, s)$ via a continuous generalization to the denoising score matching objective (Song et al. 2020b):

$$\begin{aligned}\boldsymbol{\vartheta}^* = \arg\min_{\boldsymbol{\vartheta}} \mathbb{E}_s \Big\{ \lambda(s) \mathbb{E}_{\boldsymbol{x}(0)} \mathbb{E}_{\boldsymbol{x}(s)|\boldsymbol{x}(0)} \Big[\big\| \boldsymbol{s}_{\boldsymbol{\vartheta}}(\boldsymbol{x}(s), s) \\ - \nabla_{\boldsymbol{x}(s)} \log p_{0s}(\boldsymbol{x}(s) \mid \boldsymbol{x}(0)) \big\|_2^2 \Big] \Big\}.\end{aligned} \tag{4}$$

Here $\lambda : [0, S] \to \mathbb{R}^+$ is a positive weighting function, s is uniformly sampled over $[0, T]$, $\boldsymbol{x}(0) \sim p_0(\boldsymbol{x})$ and $\boldsymbol{x}(s) \sim p_{0s}(\boldsymbol{x}(s) \mid \boldsymbol{x}(0))$. With sufficient data and

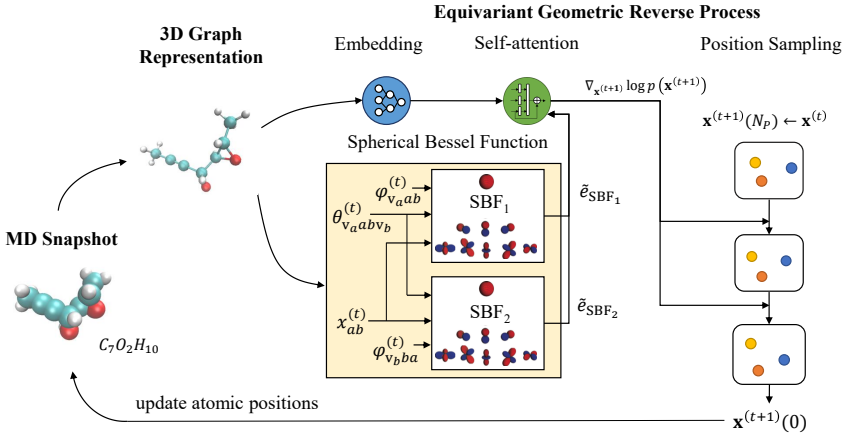

Figure 1: The overall procedure of our DIFFMD. Starting from the conformation of the last time step, atomic locations are sequentially updated with the gradient information from the score network.

model capacity, score matching ensures that the optimal solution to Eq. 4, denoted by $\boldsymbol{s}_{\boldsymbol{\vartheta}^*}(\boldsymbol{x}, s)$, equals $\nabla_{\boldsymbol{x}} \log p_s(\boldsymbol{x})$ for almost all $\boldsymbol{x}$ and s. We can typically choose $\lambda \propto 1/\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}(s)} \log p_{0s}(\boldsymbol{x}(s) \mid \boldsymbol{x}(0))\right\|_2^2\right]$ (Song et al. 2020b).

DIFFMD

Model Overview

Most prior DLMD studies such as Zhang et al. (2018) rely on the potential energy U as the intermediate variable to acquire atomic forces and update positions, which demands an additional backpropagation calculation and significantly increases the computational costs. Some recent work starts to abandon the two-stage manner and choose the atom-level force $\boldsymbol{F}$ as the prediction target of deep networks (Park et al. 2021). However, they all rely on the integrator from external computational tools to renew the positions in accordance with pre-calculated energy or forces. None embraces a straightforward paradigm to immediately forecast the 3D coordinates in a microscopic system concurrently based on previously available timeframes, i.e., $p\left(\boldsymbol{x}^{(t+1)} \mid \{\mathcal{M}^{(i)}\}_{i=0}^{t}\right)$. To bridge this gap, we seek to generate trajectories without any transitional integrator.

Several MD simulation frameworks assume the Markov property on biomolecular conformational dynamics (Chodera and Noé 2014; Malmstrom et al. 2014) for ease of representation, i.e., $p\left(\boldsymbol{x}^{(t+1)} \mid \{\mathcal{M}^{(i)}\}_{i=0}^{t}\right) = p\left(\boldsymbol{x}^{(t+1)} \mid \mathcal{M}^{(t)}\right)$. We also hold this assumption and aim to estimate the gradient field of the log density of atomic positions at each timeframe, i.e. $\nabla_{\boldsymbol{x}^{(t+1)}} \log p\left(\boldsymbol{x}^{(t+1)}\right)$. In this setting, we design a score network based on the Transformer architecture to learn the scores of the position distribution, i.e., $\boldsymbol{s}_{\boldsymbol{\vartheta}}\left(\mathcal{M}^{(t+1)}\right) = \nabla_{\boldsymbol{x}^{(t+1)}} \log p\left(\boldsymbol{x}^{(t+1)}\right)$. During the inference period, we regard the conformation of the previous frame $\mathcal{M}^t$ as the prior distribution, from which $\boldsymbol{x}^{t+1}$ is sampled. Note that $\boldsymbol{s}_{\boldsymbol{\vartheta}}\left(\mathcal{M}^{(t+1)}\right) \in \mathbb{R}^N$, we formulate it as a node regression problem. The whole procedure of DIFFMD is depicted in Fig. 1.

Score-based Generative Models for MD

The motivation for our extending the denoising diffusion models to MD simulations is their resemblance to the enhanced sampling mechanism. Inspired by non-equilibrium statistical physics, these models first systematically and slowly destroy structures in distribution through an iterative forward diffusion process and then reverse it, similar to the behavior of perturbing the free energy in the system and striving to minimize the overall energy.

Perturbing data conditionally with SDEs. Our goal is to construct a diffusion process $\left\{\boldsymbol{x}^{(t+1)}(s)\right\}_{s=0}^{S}$ indexed by a continuous time variable $s \in [0, S]$, such that $\boldsymbol{x}^{(t+1)}(0) \sim p_0$ and $\boldsymbol{x}^{(t+1)}(S) \sim p_S$. There, p_0 and p_S are the data distribution and the prior distribution of atomic positions respectively, as Equation 2.

How to incorporate noise remains critical to the success of the generation, which ensures the resulting distribution does not collapse to a low dimensional manifold (Song and Ermon 2019). Conventionally, p_S is an unstructured prior distribution, such as a Gaussian distribution with fixed mean and variance (Song et al. 2020b), which is uninformative for p_0. This construction of p_S improves the sample variety for image generation (Brock, Donahue, and Simonyan 2018) but may not work well for MD. One reason is corrupting molecular conformations unconditionally would trigger severe turbulence to the microscopic system; besides, it ignores the fact that molecular conformations of neighboring frames $\mathcal{M}^{(t)}$ and $\mathcal{M}^{(t+1)}$ are close to each other and their divergence is dependent on the status of the former one. Therefore, it is necessary to formulate p_S with the prior knowledge of $\mathcal{M}^{(t)}$.

To be explicit, the noise does not constantly grow along with s, but depends on prior states. This strategy aligns with the Gaussian accelerated MD (GaMD) mechanism (details are in Appendix) and serve as a more practical way to inject turbulence into p_0. Driven by the foregoing analysis, we

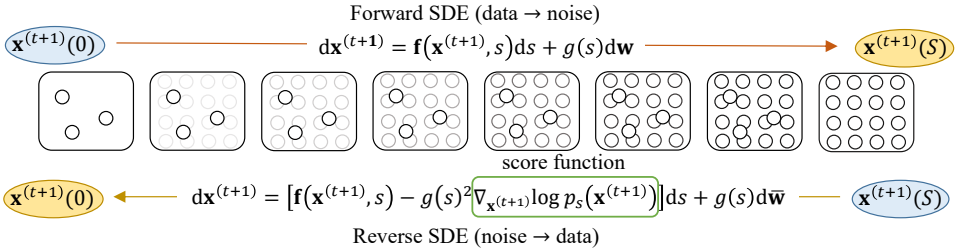

Figure 2: Solving a reverse-time SDE yields a score-based model to predict positions. The cycles indicate the atomic locations, and the darkness represents the noise strengths.

introduce a conditional noise in compliance with the accelerations at prior frames and choose the following SDE:

$$d\boldsymbol{x}^{(t+1)} = \sigma^s_{\boldsymbol{a}^{(t)}} d\boldsymbol{w}, s \in [0, 1], \tag{5}$$

where the noise term $\sigma^s_{\boldsymbol{a}^{(t)}}$ is dynamically adjusted as:

$$\sigma^s_{\boldsymbol{a}^{(t)}} = \sigma_s \eta_\sigma \left(\left\| \boldsymbol{a}^{(t-1)} \right\|_2^2 - \bar{a} \right)^2, \left\| \boldsymbol{a}^{(t-1)} \right\|_2^2 < \bar{a}, \tag{6}$$

here η_σ is the harmonic acceleration constant and $\bar{a}$ represents the acceleration threshold. Once the system has a slow variation trend of motion (i.e., the systematical energy is low), it will be supplied with a large level of noise and verse vice. Thus, the conditional noise is inversely proportional to $\boldsymbol{a}^{(t-1)}$ and inherits the merits of enhanced sampling.

Generating samples through reversing the SDE. Following the reverse-time SDE (Anderson 1982), samples of the next timeframe $\boldsymbol{x}^{(t+1)}$ can be attained by reversing the diffusion process as:

$$\begin{aligned} \mathrm{d}\boldsymbol{x}^{(t+1)} &= g(s)\mathrm{d}\overline{\boldsymbol{w}} \\ &+ \left[f\left(\boldsymbol{x}^{(t+1)}, s\right) - g(s)^2 \nabla_{\boldsymbol{x}^{(t+1)}} \log p_s\left(\boldsymbol{x}^{(t+1)}\right) \right] \mathrm{d}s. \end{aligned} \tag{7}$$

Once the score of each marginal distribution, $\nabla_{\boldsymbol{x}^{(t+1)}} \log p_s\left(\boldsymbol{x}^{(t+1)}\right)$, is known for all s, we can simulate the reverse diffusion process to sample from p_0. The workflow is summarized in Fig. 2.

Estimating scores for the SDE. Intuitively, the optimal parameters $\boldsymbol{\vartheta}^*$ of the conditional score network $\boldsymbol{s}_{\boldsymbol{\vartheta}}\left(\tilde{\mathcal{M}}^{(t+1)}\right)$ can be trained directly by minimizing the following formula:

$$\begin{aligned} \mathbb{E}_s \Bigg\{ \lambda(s) \mathbb{E}_{\boldsymbol{x}^{(t+1)}(0)} \mathbb{E}_{\boldsymbol{x}^{(t+1)}(s)|\boldsymbol{x}^{(t+1)}(0)} \Bigg[\Bigg\| \boldsymbol{s}_{\boldsymbol{\vartheta}}\left(\tilde{\mathcal{M}}^{(t+1)}(s), s\right) \\ - \nabla_{\boldsymbol{x}^{(t+1)}(s)} \log p_{0s}\left(\boldsymbol{x}^{(t+1)}(s) \mid \boldsymbol{x}^{(t+1)}(0)\right) \Bigg\|_2^2 \Bigg] \Bigg\}. \end{aligned} \tag{8}$$

Here $\boldsymbol{x}^{(t+1)}(0) \sim p_0\left(\boldsymbol{x}^{(t+1)}\right)$ and $\boldsymbol{x}^{(t+1)}(s) \sim p_{0s}\left(\boldsymbol{x}^{(t+1)}(s) \mid \boldsymbol{x}^{(t+1)}(0)\right)$. $\tilde{\mathcal{M}}^{(t+1)}$ stands for the disturbed conformation with the noised geometric position $\tilde{\boldsymbol{x}}^{(t+1)}$. Notably, other score matching objectives, such as sliced score matching (Song et al. 2020a) and finite-difference score matching (Pang et al. 2020) can also be applied here rather than denoising score matching in Eq. 8.

In order to efficiently solve Eq. 8, it is required to know the transition kernel $p_{0s}\left(\boldsymbol{x}^{(t+1)}(s) \mid \boldsymbol{x}^{(t+1)}(0)\right)$. When $f(., s)$ is affine, this transition kernel is always a Gaussian distribution, where its mean and variance are in closed forms by means of standard techniques (Särkkä and Solin 2019):

$$\begin{aligned} &p_{0s}\left(\boldsymbol{x}^{(t+1)}(s) \mid \boldsymbol{x}^{(t+1)}(0)\right) = \\ &\mathcal{N}\left(\boldsymbol{x}^{(t+1)}(s); \boldsymbol{x}^{(t+1)}(0), \frac{1}{2\log \sigma_{\boldsymbol{a}^{(t)}}} \left(\sigma^{2s}_{\boldsymbol{a}^{(t)}} - 1\right) \boldsymbol{I}\right). \end{aligned} \tag{9}$$

Equivariant Geometric Score Network

Equivariance is a ubiquitous symmetry, which complies with the fact that physical laws hold regardless of the coordinate system. It has shown efficacy to integrate such inductive bias into model parameterization for modeling 3D geometry (Köhler, Klein, and Noé 2020; Klicpera, Becker, and Günnemann 2021). Hence, we consider building the score network $\boldsymbol{s}_{\boldsymbol{\vartheta}}$ equivariant to rotation and translation transformations.

Existing equivariant models for molecular representations are static rather than kinetic. In contrast, along MD trajectories each atom has a velocity and a corresponding orientation. To be specific, for some pair of atoms (a, b), they formulate two intersecting planes (see Fig. 3) with respective velocities $\left(\boldsymbol{v}^{(t)}_a, \boldsymbol{v}^{(t)}_b\right)$. We denote the angles between velocities and the connecting line of two atoms by $\varphi^{(t)}_{\boldsymbol{v}_a ab} = \angle \hat{\boldsymbol{v}}^{(t)}_a \hat{\boldsymbol{x}}^{(t)}_{\boldsymbol{v}_b ba}$ and $\varphi^{(t)}_{\boldsymbol{v}_b ba} = \angle \hat{\boldsymbol{v}}^{(t)}_b \hat{\boldsymbol{x}}^{(t)}_{ba}$. We denote the dihedral angle between two half-phases as $\theta^{(t)}_{\boldsymbol{v}_a ab \boldsymbol{v}_b} = \angle \hat{\boldsymbol{v}}^{(t)}_a \hat{\boldsymbol{v}}^{(t)}_b \perp \hat{\boldsymbol{x}}^{(t)}_{ab}$. These three angles contain pivotal geometric information for predicting pairwise interactions as well as their future positions. It is necessary to incorporate them into our geometric modeling. Unfortunately, the directions and velocities of atomic motion, uniquely owned by dynamic systems, are seldom concerned by prior models.

To this end, we draw inspiration from Equivariant Graph Neural Networks (EGNN) (Satorras, Hoogeboom, and Welling 2021), GemNet (Klicpera, Becker, and Günnemann 2021), and Molformer (Wu et al. 2021), and propose an equivariant geometric Transformer (EGT) as $\boldsymbol{s}_{\boldsymbol{\vartheta}}$. Our

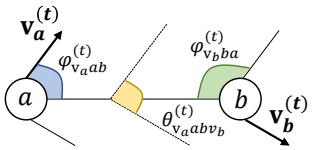

Figure 3: Angles leveraged in our EGT, including two intersection angles, $\varphi^{(t)}_{\boldsymbol{v}_a ab}$ and $\varphi^{(t)}_{\boldsymbol{v}_b ba}$,and one dihedral angle, $\theta^{(t)}_{\boldsymbol{v}_a ab \boldsymbol{v}_b}$.

EGT is roto-translation equivariant and leverages directional information. The l-th equivariant geometric layer (EGL) takes the set of atomic coordinates $\boldsymbol{x}^{(t),l}$, velocities $\boldsymbol{v}^{(t),l}$, and features $\boldsymbol{h}^{(t),l}$ as input, and outputs a transformation on $\boldsymbol{x}^{(t),l+1}$, $\boldsymbol{v}^{(t),l+1}$, and $\boldsymbol{h}^{(t),l+1}$. Concisely, $\boldsymbol{x}^{(t),l+1}, \boldsymbol{v}^{(t),l+1}, \boldsymbol{h}^{(t),l+1} = \text{EGL}\left(\boldsymbol{x}^{(t),l}, \boldsymbol{v}^{(t),l}, \boldsymbol{h}^{(t),l}\right)$.

We first calculate the spherical Fourier-Bessel bases to integrate all available geometric information:

$$\tilde{e}^{l}_{\text{SBF}_1,omn}\left(x^{(t),l}_{ab}, \varphi^{(t),l}_{\boldsymbol{v}_a ab}, \theta^{(t),l}_{\boldsymbol{v}_a ab \boldsymbol{v}_b}\right) = \sqrt{\frac{2}{c^3_{\text{int}} j^2_{o+1}(z_{on})}} j_o\left(\frac{z_{on}}{c_{\text{int}}} x^{(t),l}_{ab}\right) Y_{om}\left(\varphi^{(t),l}_{\boldsymbol{v}_a ab}, \theta^{(t),l}_{\boldsymbol{v}_a ab \boldsymbol{v}_b}\right), \tag{10}$$

$$\tilde{e}^{l}_{\text{SBF}_2,omn}\left(x^{(t),l}_{ab}, \varphi^{(t),l}_{\boldsymbol{v}_b ba}, \theta^{(t),l}_{\boldsymbol{v}_a ab \boldsymbol{v}_b}\right) = \sqrt{\frac{2}{c^3_{\text{int}} j^2_{o+1}(z_{on})}} j_o\left(\frac{z_{on}}{c_{\text{int}}} x^{(t),l}_{ab}\right) Y_{om}\left(\varphi^{(t),l}_{\boldsymbol{v}_b ba}, \theta^{(t),l}_{\boldsymbol{v}_a ab \boldsymbol{v}_b}\right), \tag{11}$$

where $o \in [N_{\text{CBF}}]$, $n \in [N_{\text{RBF}}]$, and $m \in [N_{\text{SBF}}]$ control the degree, root, and order of the radial basis functions, respectively. c_{int} is the interaction cutoff. j_o is the spherical Bessel functions. z_{on} is the n-th root of the o-degree Bessel functions. Y_{om} is the real spherical harmonics with degree o and order m. Remarkably, 3D spherical Fourier-Bessel representations including $\tilde{\boldsymbol{e}}_{\text{SBF}_1}$ and $\tilde{\boldsymbol{e}}_{\text{SBF}_2}$ enjoy the roto-translation invariant property due to their exploitation of the relative distance as well as the invariant angles. Then those directional vectors are fed into EGL as:

$$\boldsymbol{q}_i = \left[f_q\left(\boldsymbol{h}^{(t),l}_i\right) \oplus \tilde{\boldsymbol{e}}^{l}_{\text{SBF}_1}\right] \boldsymbol{W}_{\text{SBF}_1}, \tag{12}$$

$$\boldsymbol{k}_i = \left[f_k\left(\boldsymbol{h}^{(t),l}_i\right) \oplus \tilde{\boldsymbol{e}}^{l}_{\text{SBF}_2}\right] \boldsymbol{W}_{\text{SBF}_2}, \tag{13}$$

$$\boldsymbol{m}_i = f_m\left(\boldsymbol{h}^{(t),l}_i\right),\ a_{ij} = \boldsymbol{q}_i \boldsymbol{k}^T_j / \sqrt{\psi_{\text{att}}}, \tag{14}$$

$$\boldsymbol{v}^{(t),l+1}_i = f_v\left(\boldsymbol{h}^{(t),l}_i\right) \boldsymbol{v}^{(t),l}_i + \sum_{j=1}^{N} \phi\left(a_{ij}\right) \boldsymbol{x}^{(t),l}_{ij}, \tag{15}$$

$$\boldsymbol{x}^{(t),l+1}_i = \boldsymbol{x}^{(t),l}_i + \frac{1}{L} \boldsymbol{v}^{(t),l+1}_i, \tag{16}$$

$$\boldsymbol{h}^{(t),l+1}_i = f_h\left(\sum_{j=1}^{N} \phi\left(a_{ij}\right) \boldsymbol{m}_j\right). \tag{17}$$

Here $\oplus$ denotes concatenation and L is the number of total layers in EGT. f_q, f_k and f_m are linear transformations. f_v and f_h are velocity and feature operations, which are commonly approximately by multi-layer perceptrons (MLPs). $\boldsymbol{q}_i$, $\boldsymbol{k}_i$ and $\boldsymbol{m}_i$ are respectively the query, key, and value vectors with the same dimension ψ_{att}. The weight matrix $\boldsymbol{W}_{\text{SBF}_1}$ and $\boldsymbol{W}_{\text{SBF}_2}$ are learnable, transferring dimensions of the concatenated vectors back to ψ_{att}. a_{ij} is the attention that the token i pays to the token j. ϕ denotes the *Softmax* function. Finally, $\boldsymbol{x}^{(t),L}$ at the last layer immediately draw the gradient field of locations, i.e., $\nabla_{\boldsymbol{x}^{(t)}} \log p\left(\boldsymbol{x}^{(t)}\right)$.

Note that EGL breaks down the coordinate update into two stages. First we compute the velocity $\boldsymbol{v}^{(t),l+1}_i$, and then leverage it to update the position $\boldsymbol{x}^{(t),l}_i$. The initial velocity $\boldsymbol{v}^{(t),l}_i$ is scaled by $f_v : \mathbb{R}^{\psi_h} \to \mathbb{R}$ that maps the feature embedding $\boldsymbol{h}^{(t),l}_i$ to a scalar value. After that, the velocity of each atom $\boldsymbol{v}^{(t),l}_i$ is updated as a vector field in a radial direction. In other words, $\boldsymbol{v}^{(t),l}_i$ is renewed by the weighted sum of all relative differences $\left\{\boldsymbol{x}^{(t),l}_{ij}\right\}_{j=1}^{N}$. The weights of this sum are provided as the attention score $\{a_{ij}\}_{j=1}^{N}$. Meanwhile, those attention scores are used to gain the new feature $\boldsymbol{h}^{(t),l+1}_i$.

Analysis on E(n) equivariance. We analyze the equivariance properties of our model for E(3) symmetries. That is, our model should be translation equivariant on $\boldsymbol{x}$ for any translation vector and it should also be rotation and reflection equivariant on $\boldsymbol{x}$ for any orthogonal matrix $Q \in \mathbb{R}^{n \times n}$ and any translation matrix $o \in \mathbb{R}^{n \times 3}$. More formally, our model satisfies (see proof in Appendix):

$$Q\boldsymbol{x}^{(t),l+1} + o, Q\boldsymbol{v}^{(t),l+1}, \boldsymbol{h}^{(t),l+1} = \text{EGL}\left(Q\boldsymbol{x}^{(t),l} + o, Q\boldsymbol{v}^{(t),l}, \boldsymbol{h}^{(t),l}\right). \tag{18}$$

Trajectory Sampling

After training a time-dependent score-based model $\boldsymbol{s}_{\boldsymbol{\vartheta}}$, we can exploit it to construct the reverse-time SDE and then simulate it with numerical methods to generate molecular conformations from p_0. As analyzed before, $\boldsymbol{x}^{(t)}$ and $\boldsymbol{x}^{(t+1)}$ are heavily correlated and their divergence is minor. Based on this relationship, instead of using some Gaussian distributions (Song et al. 2020b), we leverage $\boldsymbol{x}^{(t)}$ as a replacement to approximate the unknown prior distribution $p_S\left(\boldsymbol{x}^{(t+1)}\right)$. That is, we regard $\boldsymbol{x}^{(t)}$ as the perturbed version of $\boldsymbol{x}^{(t+1)}$ and seize it as the starting point of our trajectory sampling process.

Numerical solvers provide approximate computation for SDEs. Many general-purpose numerical methods, such as Euler-Maruyama and stochastic Runge-Kutta methods, apply to the reverse-time SDE for sample generation. In addition to SDE solvers, we can also employ score-based Markov Chain Monte Carlo (MCMC) approaches such as Langevin MCMC or Hamiltonian Monte Carlo to sample from directly, and correct the solution of a numerical SDE

Algorithm 1: Sampling Algorithm with Predictor-Corrector.

Require: N_P: Number of discretization steps for the reverse-time SDE.
Require: N_C: Number of corrector steps.
  Initialize $\boldsymbol{x}^{(t+1)}(N_P) \leftarrow \boldsymbol{x}^{(t)}$
  for $i = N_P - 1$ to 0 **do**
    $\boldsymbol{x}^{(t+1)}(i) \leftarrow \text{Predictor}\left(\boldsymbol{x}^{(t+1)}(i+1)\right)$
    for $j = 1$ to $N_C - 1$ **do**
      $\boldsymbol{x}^{(t+1)}(i) \leftarrow \text{Corrector}\left(\boldsymbol{x}^{(t+1)}(i)\right)$
    end for
  end for
  return $\boldsymbol{x}^{(t+1)}$

solver. Readers are recommended to refer to Song et al. (2020b) for more details. We provide the pseudo-code of the whole sampling process in Algorithm 1.

Experiments

To verify the effectiveness of our DIFFMD, we construct the following two tasks and empirically evaluate it:

Short-term-to-long-term (S2L) Trajectory Generation. In this task setting, models are first trained on some short-term trajectories and are required to produce long-term trajectories of the same molecule given the starting conformation $\boldsymbol{x}^{(t_0)}$ as $p\left(\boldsymbol{x}^{(t_n)}, ..., \boldsymbol{x}^{(t_1)} | \boldsymbol{x}^{(t_0)}\right)$. This extrapolation over time aims to examine the model's capacity of generalization from the temporal view.

One-to-others (O2O) Trajectory Generation. In the O2O task, models are trained on the entire trajectories of some molecules and examined on other molecules from scratch. This evaluates model's eligibility to generalize to conformations of different molecules, namely, the discrepancy with respect to different molecular types.

Experiment Setup

Evaluation metric. We adopt the accumulative root-mean-square-error (ARMSE) of all snapshots at a given n-step time period $\{t_i\}_{i=1}^n$ as the evaluation metric. ARMSE evaluates the generated conformations as: $\text{ARMSE} = \left(\frac{1}{n}\sum_{i=t_1}^{t_n}\left\|\tilde{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\right\|^2\right)^{\frac{1}{2}}$, which is roto-translational invariant.

Baselines. We compare DIFFMD with several state-of-the-art methods for the MD trajectory prediction. Specifically, **Tensor Field Network (TFN)** (Thomas et al. 2018) adopts filters built from spherical harmonics to achieve equivariance. **Radial Field (RF)** is a GNN drawn from Equivariant Flows (Köhler, Klein, and Noé 2019). **SE(3)-Transformer** (Fuchs et al. 2020) is a equivariant variant of the self-attention module for 3D point-clouds. **EGNN** (Satorras, Hoogeboom, and Welling 2021) learns GNNs equivariant to rotations, translations, reflections and permutations. **GMN** (Huang et al. 2022) resorts to generalized coordinates to impose geometrical constraints on graphs. **SCFNN** (Gao and Remsing 2022) is a self-consistent field NN for learning the long-range response of molecular systems. The full experimental details are elaborated in Appendix.

Short-term-to-long-term Trajectory Generation

Data. MD17 (Chmiela et al. 2017) [1] contains trajectories of eight thermalized molecules, and all are calculated at a temperature of 500K and a resolution of 0.5 femtosecond (ft). We use the first 20K frame pairs as the training set and split the next 20K frame pairs equally into validation and test sets. Unfortunately, MD17 does not include velocities of particles, for which we use $\boldsymbol{v}^{(t)} = \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}$ as a substitution, similarly to GMN.

Results. Table 1 documents the performance of baselines and our DIFFMD in S2L, where the best performance is marked bold and the second best is underlined for clear comparison. Note that floating overflow is encountered by RF (denoted as NA). For all eight organic molecules, DIFFMD achieves the lowest ARMSEs. Moreover, different organic molecules perform in different manners during MD. Particularly, benzene moves most actively than other molecules, which leads to the highest prediction errors.

One-to-others Trajectory Generation

Data. $C_7O_2H_{10}$ (Brockherde et al. 2017) [1] is a dataset that consists of the trajectories of 113 randomly selected $C_7O_2H_{10}$ isomers, which are calculated at a temperature of 100K and resolution of 1 fs using density functional theory with the PBE exchange-correlation potential. We select the top-5 isomers that have the largest ARMSEs out of 113 samples, using $\boldsymbol{x}^{(t_0)}$ as the prediction for all the subsequent timeframes, as the validation targets and take the rest as the training set. Same as the MD17 case, we compute the distance vector between neighboring frames as the velocities.

Results. Table **??** reports ARMSE of baselines and our DIFFMD on the five isomers from $C_7O_2H_{10}$. DIFFMD exceeds all baselines with a large margin for all target molecules. We plot snapshots at different timeframes in Appendix.

A closer inspection of the generated trajectories shows that several baselines have worse generation quality because their conformations are not geometrically and biologically constrained. On the contrary, generated conformations by models like EGNN and GMN are geometrically legal, but their variations are minute. Interestingly, we discover that conformations generated by SCFNN remain unchanged after a few timeframes, which indicates the network is stuck in a fixed point.

Related Work

Molecular Dynamics with Deep Learning Recently, various DL models have become easy-to-use tools for fasci-

[1] Both MD17 and $C_7O_2H_{10}$ datasets are available at http://quantum-machine.org/datasets/

Methods	Aspirin	Benzene	Ethanol	Malonaldehyde	Naphthalene	Salicylic	Toluene	Uracil
TFN	NA	NA	NA	NA	NA	NA	NA	NA
RF	3.707	19.724	5.963	18.532	13.791	2.071	4.052	2.382
SE(3)-Tr.	0.813	2.415	0.678	1.183	1.834	1.230	1.312	0.691
EGNN	0.868	2.518	0.719	0.889	0.484	0.632	1.034	0.464
GMN	0.814	2.528	0.751	0.880	0.832	0.895	1.018	0.494
SCFNN	1.151	2.832	1.084	1.096	0.923	0.918	1.229	0.857
DIFFMD	**0.648**	**2.365**	**0.637**	**0.784**	**0.298**	**0.471**	**0.820**	**0.393**
Relative Impro.	20.2%	2.1%	5.9%	10.9%	38.4%	25.4%	19.5%	15.4%

Table 1: Extrapolation performance on MD17. Note the extrapolation errors for TFN are not available (NA) due to the floating number overflow.

Methods	ISO_1004	ISO_2134	ISO_2126	ISO_3001	ISO_1007
TFN	7.390	10.990	10.412	4.697	10.677
RF	4.772	4.364	21.576	9.077	11.049
SE(3)-Tr.	5.253	6.186	4.334	5.304	7.514
EGNN	1.142	0.578	0.928	1.017	1.035
GMN	1.205	0.363	0.998	1.053	1.154
SCFNN	1.781	1.693	1.785	2.842	2.264
DIFFMD	**1.127**	**0.278**	**0.919**	**0.837**	**0.878**
Relative Impro.	1.2%	23.4%	9.7%	9.8%	15.1%

Table 2: Performance on the five isomers in $C_7O_2H_{10}$.

nating MD with *ab initio* accuracy. Behler-Parrinello network (Behler and Parrinello 2007) is one of the first models to learn potential surfaces from MD data. After that, DeepPotential net (Han et al. 2017) is further developed by extending to more advanced functions involving two neighbors. While DTNN (Schütt et al. 2017) and SchNet (Schütt et al. 2018) achieve highly competitive prediction performance across the chemical compound space and the configuration space in order to simulate MD (Noé et al. 2020). However, they still follow the routine of multi-stage simulations and rely on forces or energy as the prediction target. Huang et al. (2022) proposes an end-to-end GMN to characterize constrained systems of interacting objects, where molecules are defined as a set of rigidly connected particles with sticks and hinges. Also, their experiments fail to be realistic and the constraint strongly violates the nature of MD, since no distance between any pair of atoms are fixed.

Conformation Generation. Researchers are increasingly interested in conformation generation. Some works start from 2D molecular graphs to gain their 3D structures via bi-level programming (Xu et al. 2021b) and continuous flows (Xu et al. 2021a). Some others concentrate on the inverse design to create the conformations of drug-like molecules or crystals with desired properties (Noh et al. 2019). Recently, Gao and Remsing (2022) propose an SCFNN that perturbs positions of the Wannier function centers induced by external electric fields. Latterly, diffusion models become a favored choice in conformation generation (Shi et al. 2021). Xu et al. (2022) introduce a GeoDiff by progressively injecting and eliminating small noises. However, its perturbations evolve over discrete times. A better approach would be to express dynamics as a set of differential equations since time is actually continuous. Furthermore, these studies leverage diffusion models in recovering conformations from molecular graphs instead of generating sequential conformations. We fill in the gap by applying them to yield MD trajectories.

Conclusion

We propose DIFFMD, a novel principle to sequentially generate molecular conformations in MD simulations. DIFFMD marries denoising diffusion models with an equivariant geometric Transformer, which enables self-attention to leverage directional information in addition to the interatomic distances. Extensive experiments over multiple tasks demonstrate that DIFFMD is superior to existing state-of-the-art models. This research may shed light on the acceleration of new drugs and material discovery.

References

Anderson, B. D. 1982. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3): 313–326.

Behler, J.; and Parrinello, M. 2007. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14): 146401.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; and Müller, K.-R. 2017. Bypassing the Kohn-Sham equations with machine learning. *Nature communications*, 8(1): 1–10.

Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; and Müller, K.-R. 2017. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5): e1603015.

Chodera, J. D.; and Noé, F. 2014. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25: 135–144.

Cranmer, M.; Greydanus, S.; Hoyer, S.; Battaglia, P.; Spergel, D.; and Ho, S. 2020. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*.

De Vivo, M.; Masetti, M.; Bottegoni, G.; and Cavalli, A. 2016. Role of molecular dynamics and related methods

in drug discovery. *Journal of medicinal chemistry*, 59(9): 4035–4061.

Dror, R. O.; Arlow, D. H.; Maragakis, P.; Mildorf, T. J.; Pan, A. C.; Xu, H.; Borhani, D. W.; and Shaw, D. E. 2011. Activation mechanism of the β2-adrenergic receptor. *Proceedings of the National Academy of Sciences*, 108(46): 18684–18689.

Fuchs, F.; Worrall, D.; Fischer, V.; and Welling, M. 2020. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33: 1970–1981.

Gao, A.; and Remsing, R. C. 2022. Self-consistent determination of long-range electrostatics in neural network potentials. *Nature Communications*, 13(1): 1–11.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.

Han, J.; Zhang, L.; Car, R.; et al. 2017. Deep potential: A general representation of a many-body potential energy surface. *arXiv preprint arXiv:1707.01478*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, W.; Han, J.; Rong, Y.; Xu, T.; Sun, F.; and Huang, J. 2022. Equivariant Graph Mechanics Networks with Constraints. In *International Conference on Learning Representations*.

Jia, W.; Wang, H.; Chen, M.; Lu, D.; Lin, L.; Car, R.; Weinan, E.; and Zhang, L. 2020. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. In *SC20: International conference for high performance computing, networking, storage and analysis*, 1–14. IEEE.

Klicpera, J.; Becker, F.; and Günnemann, S. 2021. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34.

Klicpera, J.; Groß, J.; and Günnemann, S. 2020. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*.

Köhler, J.; Klein, L.; and Noé, F. 2019. Equivariant flows: sampling configurations for multi-body systems with symmetric energies. *arXiv preprint arXiv:1910.00753*.

Köhler, J.; Klein, L.; and Noé, F. 2020. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International Conference on Machine Learning*, 5361–5370. PMLR.

Liao, Q. 2020. Enhanced sampling and free energy calculations for protein simulations. In *Progress in molecular biology and translational science*, volume 170, 177–213. Elsevier.

Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; and Shaw, D. E. 2011. How fast-folding proteins fold. *Science*, 334(6055): 517–520.

Luo, R.; Zhang, Q.; Yang, Y.; and Wang, J. 2020. Replica-exchange Nos\'e-Hoover dynamics for Bayesian learning on large datasets. *Advances in Neural Information Processing Systems*, 33: 17874–17885.

Malmstrom, R. D.; Lee, C. T.; Van Wart, A. T.; and Amaro, R. E. 2014. Application of molecular-dynamics based markov state models to functional proteins. *Journal of chemical theory and computation*, 10(7): 2648–2657.

Miao, Y.; Feher, V. A.; and McCammon, J. A. 2015. Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. *Journal of chemical theory and computation*, 11(8): 3584–3595.

Noé, F.; Tkatchenko, A.; Müller, K.-R.; and Clementi, C. 2020. Machine learning for molecular simulation. *Annual review of physical chemistry*, 71: 361–390.

Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; and Jung, Y. 2019. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5): 1370–1384.

Pang, T.; Xu, K.; Li, C.; Song, Y.; Ermon, S.; and Zhu, J. 2020. Efficient learning of generative models via finite-difference score matching. *Advances in Neural Information Processing Systems*, 33: 19175–19188.

Park, C. W.; Kornbluth, M.; Vandermause, J.; Wolverton, C.; Kozinsky, B.; and Mailoa, J. P. 2021. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *npj Computational Materials*, 7(1): 1–9.

Rocchia, W.; Masetti, M.; and Cavalli, A. 2012. Enhanced sampling methods in drug design. *Physico-Chemical and Computational Approaches to Drug Discovery. The Royal Society of Chemistry*, 273–301.

Särkkä, S.; and Solin, A. 2019. *Applied stochastic differential equations*, volume 10. Cambridge University Press.

Satorras, V. G.; Hoogeboom, E.; and Welling, M. 2021. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*, 9323–9332. PMLR.

Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; and Tkatchenko, A. 2017. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1): 1–8.

Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; and Müller, K.-R. 2018. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24): 241722.

Schwantes, C. R.; McGibbon, R. T.; and Pande, V. S. 2014. Perspective: Markov models for long-timescale biomolecular dynamics. *The Journal of chemical physics*, 141(9): 09B201_1.

Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; and Shaw, D. E. 2011. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24): 9181–9183.

Shi, C.; Luo, S.; Xu, M.; and Tang, J. 2021. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, 9558–9568. PMLR.

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

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32.

Song, Y.; Garg, S.; Shi, J.; and Ermon, S. 2020a. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, 574–584. PMLR.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Suomivuori, C.-M.; Gamiz-Hernandez, A. P.; Sundholm, D.; and Kaila, V. R. 2017. Energetics and dynamics of a light-driven sodium-pumping rhodopsin. *Proceedings of the National Academy of Sciences*, 114(27): 7043–7048.

Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; and Riley, P. 2018. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, F.; Zhang, Q.; Radev, D.; Cui, J.; Zhang, W.; Xing, H.; Zhang, N.; and Chen, H. 2021. 3D-Transformer: Molecular Representation with Transformer in 3D Space. *arXiv preprint arXiv:2110.01191*.

Xu, M.; Luo, S.; Bengio, Y.; Peng, J.; and Tang, J. 2021a. Learning neural generative dynamics for molecular conformation generation. *arXiv preprint arXiv:2102.10240*.

Xu, M.; Wang, W.; Luo, S.; Shi, C.; Bengio, Y.; Gomez-Bombarelli, R.; and Tang, J. 2021b. An end-to-end framework for molecular conformation generation via bilevel programming. In *International Conference on Machine Learning*, 11537–11547. PMLR.

Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; and Tang, J. 2022. GeoDiff: a Geometric Diffusion Model for Molecular Conformation Generation. *arXiv preprint arXiv:2203.02923*.

Zhang, L.; Han, J.; Wang, H.; Car, R.; and Weinan, E. 2018. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14): 143001.