Fractional Denoising for 3D Molecular Pre-training

Shikun Feng [* 1]   Yuyan Ni [* 2]   Yanyan Lan [1]   Zhi-Ming Ma [2]   Weiying Ma [1]

Abstract

Coordinate denoising is a promising 3D molecular pre-training method, which has achieved remarkable performance in various downstream drug discovery tasks. Theoretically, the objective is equivalent to learning the force field, which is revealed helpful for downstream tasks. Nevertheless, there are two challenges for coordinate denoising to learn an effective force field, i.e. low sampling coverage and isotropic force field. The underlying reason is that molecular distributions assumed by existing denoising methods fail to capture the anisotropic characteristic of molecules. To tackle these challenges, we propose a novel hybrid noise strategy, including noises on both dihedral angel and coordinate. However, denoising such hybrid noise in a traditional way is no more equivalent to learning the force field. Through theoretical deductions, we find that the problem is caused by the dependency of the input conformation for covariance. To this end, we propose to decouple the two types of noise and design a novel fractional denoising method (Frad), which only denoises the latter coordinate part. In this way, Frad enjoys both the merits of sampling more low-energy structures and the force field equivalence. Extensive experiments show the effectiveness of Frad in molecular representation, with a new state-of-the-art on 9 out of 12 tasks of QM9 and on 7 out of 8 targets of MD17[1].

1. Introduction

Molecular representation learning is fundamental for various tasks in drug discovery, such as molecular property prediction (Schütt et al., 2018; 2021; Liu et al., 2022b), drug-drug interaction prediction (Asada et al., 2018; Rohani & Eslahchi, 2019), and de novo molecular generation (Gebauer et al., 2019; Luo & Ji, 2022). Inspired by the success of self-supervised learning in natural language processing (NLP) (Dai & Le, 2015; Devlin et al., 2018) and computer vision (CV) (Simonyan & Zisserman, 2014; Dosovitskiy et al., 2020), various molecular pre-training methods have been proposed, to tackle the lack of labeled data problem in this area. Among them, most early approaches treat molecular data as 1D SMILES strings (Wang et al., 2019; Honda et al., 2019; Chithrananda et al., 2020; Zhang et al., 2021a; Xue et al., 2021; Guo et al., 2022) or 2D graphs (Rong et al., 2020; Li et al., 2020; Zhang et al., 2021b; Li et al., 2021; Zhu et al., 2021; Wang et al., 2022b;c; Fang et al., 2022b; Lin et al., 2022), and utilize sequence-based or graph-based pre-training methods to obtain molecular representations. However, the 3D geometric structure is crucial for a molecule, since it largely determines the energy function, and thus the corresponding physical and chemical properties (Schütt et al., 2018). Therefore recently, more and more pre-training methods (Liu et al., 2021; Li et al., 2022; Zhu et al., 2022; Fang et al., 2022a; Stärk et al., 2022a) have been proposed to exploit 3D molecular data (see Appendix D).

In 3D molecular pre-training, the coordinate denoising approach (Zaidi et al., 2022; Luo et al., 2022; Zhou et al., 2022; Jiao et al., 2022; Liu et al., 2022a) is a promising one, and has achieved remarkable performance. Specifically given the equilibrium molecular structure, some independent and identical noise is added to the corresponding atomic coordinates, and the model gets trained to reconstruct the input. Compared with other self-supervised learning methods, coordinate denoising methods have the ability to capture the fine-grained 3D geometry information. More importantly, this approach enjoys a physical interpretation of learning a molecular force field (Zaidi et al., 2022).

Force field learning have been proven effective for downstream tasks. Theoretically, force field and potential energy are fundamental physical quantities that has close relation with several downstream tasks(Chmiela et al., 2017). Em-

[1]The code is released publicly at https://github.com/fengshikun/Frad

pirically, Zaidi et al. (2022); Jiao et al. (2022); Liu et al. (2022a); Luo et al. (2022) have demonstrated that learning the force field or energy will produce remarkable performance for various downstream tasks. To further validate this issue, we conduct additional experiments detailed in the Appendix B.3, where we employ the prediction of force field as the pre-training task, our results show the effectiveness of learning force field approach. Considering the equivalence between denoising and learning force field, denoising could be a powerful pretraining method for molecular representation. However, two challenges prevent the current coordinate denoising methods from learning an accurate force field.

- **Low Sampling Coverage.** In existing coordinate denoising methods, the noise level is usually set very small, to avoid generating irrational substructures, e.g. distorted aromatic rings. It is observed in experiments that if the noise level is large, the performance can decrease dramatically Zaidi et al. (2022). Similar phenomenon is also found in Appendix B.1. Though the existing noise sampling strategy can avoid unwanted rare noisy structures, the produced structures could hardly cover common structures with low energy, which can be crucial for various downstream tasks. Therefore, the existing coordinate denoising methods have limitations in learning accurate force field at other common low energy structures, except for the given equilibrium structures.

- **Isotropic Force Field.** In existing coordinate denoising methods, the noise is assumed to be with an isotropic covariance, meaning the slope of the energy function is the same in all directions around the local minimum. However, the energy function of molecule is intrinsically not isotropic in the sense that there can be both rigid and flexible parts in a molecule. As illustrated in Figure 1, the structures of rings, double bonds, and triple bonds, are usually fixed in low-energy conformations, while some single bonds can be rotated without causing radical energy changes. All these different structures are very popular in practice. Therefore, the existing methods fail to depict the anisotropic energy landscape, leading to inaccurate learned force field.

To tackle the aforementioned challenges, we propose a novel hybrid noise strategy to capture the characteristic of molecular distribution. Unlike the coordinate noise, we first introduce a Gaussian noise to the dihedral angles of the rotatable bonds, and then add a traditional noise to the coordinates of atoms. In this way, the dihedral angle noise scale could be set large to search the energy landscape, which may cover more meaningful low energy structures without generating invalid noisy structures. Under the setting of hybrid noise,

the corresponding conformation distribution will be with an anisotropic covariance. Especially, the covariance of the flexible parts is large through the perturbation of the rotatable dihedral angels with a large noise level. Whereas the covariance of the rigid parts is small, since only small levels of coordinate noise will be added to them.

Although the hybrid noise strategy well addresses the above two challenges, unlike traditional coordinate denoising, learning to directly recover the hybrid noise is no more equivalent to learning the force field. Through a meticulous mathematical deduction, we find that the bottleneck is the dependency of the input conformation in the formulation of covariance. Confronted with the difficulty of denoising the dihedral angles, we decouple the two types of noise and design a novel fractional denoising method. The main idea is adding the hybrid noise, while only denoising the latter coordinate part. We can prove that this new denoising method, namely fractional denoising with hybrid noise (Frad), is equivalent to learning an anisotropic force field, inheriting all the merits of hybrid noise.

The main contribution of this work is the introduction of a new hybrid noise strategy and the design of a novel fractional denoising method for 3D molecule pre-training. Theoretically, we prove that the new denoising method is equivalent to learning the force field with an anisotropic covariance, which captures the important characteristic of molecules. Empirically, we conduct experients by pre-training on a large dataset PCQM4Mv2 (Nakata & Shimazaki, 2017) and fine-tuning on two widely-used dataset, i.e. QM9 (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012) and MD17 (Chmiela et al., 2017). Experimental results show that our method achieves a new state-of-the-art on 9 out of 12 tasks of the QM9 and on 7 out of 8 targets of MD17, as compared with previous state-of-the-art denoising pre-training baselines and other approaches tailored for property prediction. Comprehensive ablation studies manifest the effectiveness of our design in both pre-training and fine-tuning.

2. Preliminary

In this section, we will clarify the widely-applied assumptions and notations in denoising pre-training and introduce
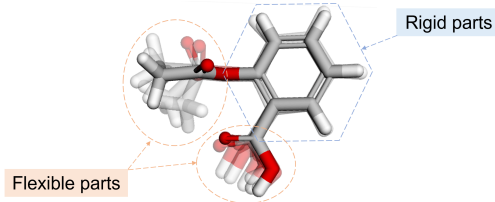

Figure 1. An illustration of the anisotropy of molecular structures. In low-energy conformations of aspirin, the structure of benzene ring and the carbon-oxygen double bonds are almost fixed, while some single bonds can rotate flexibly.

the coordinate denoising method.

Boltzmann Distribution. From the prior knowledge in statistical physics, the occurrence probability of molecular conformations is described by Boltzmann distribution (Boltzmann, 1868) $p_{physical}(\tilde{x}) \propto exp(-E_{physical}(\tilde{x}))$, where $E_{physical}(\tilde{x})$ is the (potential) energy function, $\tilde{x} \in \mathbb{R}^{3N}$ is the position of the atoms, i.e. conformation, N is the number of atoms in the molecule. More details are in Appendix C.1.

Gaussian Assumption. The goal is to learn the molecular force field $-\nabla_{\tilde{x}} E(\tilde{x})$. From the Boltzmann distribution, we have $\nabla_{\tilde{x}} \log p(\tilde{x}) = -\nabla_{\tilde{x}} E(\tilde{x})$, where $\nabla_{\tilde{x}} \log p(\tilde{x})$ is the score function of the conformation $\tilde{x}$. However both the energy function $E_{physical}$ and distribution $p_{physical}$ are unknown, and we only have access to a set of n equilibrium conformations $x_1, \cdots, x_n$ during pre-training, which are local minima of the energy and the local maxima of the probability distribution. Accordingly, the conformation distribution can be approximated by mixture of Gaussians centered at the equilibriums (Zaidi et al., 2022):

$$p_{physical}(\tilde{x}) \approx p(\tilde{x}) = \sum_{i=1}^{n} p_N(\tilde{x}|x_i)p_0(x_i), \quad (1)$$

where $p_N(\tilde{x}|x_i) \sim \mathcal{N}(x_i, \Sigma), i = 1, \cdots, n$ are Gaussian distributions, n is the number of equilibriums, $\tilde{x} \in \mathbb{R}^{3N}$ is any conformation of the molecule, p_0 is the probability of the equilibriums. Then the approximate energy function is $E(\tilde{x}) \approx E_{physical}(\tilde{x})$, which satisfies $p(\tilde{x}) \propto exp(-E(\tilde{x}))$. The Gaussian mixture degenerates into Gaussian distribution when the equilibrium is unique, and this is the case in our pre-training dataset. It is worth to note that the existing methods adopt Gaussian distribution with isotropic diagonal covariance $\Sigma = \tau^2 I_{3N}$, leading to isotropic quadratic energy function. However, it is not the case in the real world. This is exactly one of our motivations to propose a new method in section 3 to provide an anisotropic covariance matrix that better fit $p_{physical}$.

Molecular Force Field Learning. It can be proved that denoising is an equivalent optimizing objective to learning the approximate force field with the assumptions above.

$$\begin{aligned} & E_{p(\tilde{x})}\|GNN_\theta(\tilde{x}) - (-\nabla_{\tilde{x}} E(\tilde{x}))\|^2 \quad &\text{(a)} \\ =& E_{p(\tilde{x})}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x})\|^2 \quad &\text{(b)} \\ =& E_{p(\tilde{x}|x_i)p(x_i)}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}|x_i)\|^2 + T \quad &\text{(c)} \\ =& E_{p(\tilde{x}|x_i)p(x_i)}\|GNN_\theta(\tilde{x}) - \frac{x_i - \tilde{x}}{\tau^2})\|^2 + T \quad &\text{(d)}, \end{aligned} \quad (2)$$

where $GNN_\theta(\tilde{x})$ denotes a graph neural network with parameters θ which takes conformation $\tilde{x}$ as an input and returns node-level noise predictions, T is constant independent of θ, $-\nabla_x E(\tilde{x})$ is referred to as the molecular force field, indicating the force on each atom. In (2), the first equation uses the Boltzmann distribution, the second equation is proved in Vincent (2011) and Proposition

A.8 that training a neural network to estimate the score function (b) is equivalent to perturbing the data with a pre-specified noise $p(\tilde{x}|x_i)$ and training a neural network to estimate the conditional score function (c). The first two equation holds for any distribution $p(\tilde{x})$, while the last equation employs the Gaussian assumption with an isotropic diagonal variance $\Sigma = \tau^2 I_{3N}$. Since coefficients $-\frac{1}{\tau^2}$ do not rely on the input $\tilde{x}$, it can be absorbed into GNN_θ (Zaidi et al., 2022). So we conclude that typical denoising loss and force field fitting loss are equivalent, i.e. $\min_\theta E_{p(\tilde{x}|x_i)p(x_i)}\|GNN_\theta(\tilde{x}) - (\tilde{x} - x_i)\|^2 \simeq \min_\theta E_{p(\tilde{x})}\|GNN_\theta(\tilde{x}) - (-\nabla_{\tilde{x}} E(\tilde{x}))\|^2$, where $\simeq$ denotes equivalent optimization objectives for GNN. This proof helps to comprehend the content in section 3.2 and 3.3. More details are in Appendix C.2.

3. Method

In this section, we clarify the two challenges faced by traditional coordinate denoising in detail and elaborate how we tackle them by designing dihedral angle noise and hybrid noise in section 3.1. Then in section 3.2, we provide a mathematical description for the two types of noise and explain that the force field interpretation does not hold for directly denoising the hybrid noise. Finally, to eliminate this limitation, a new kind of denoising task is proposed in section 3.3.

3.1. Hybrid Noise

In fact, the fundamental cause of the two challenges is the inadequate distribution assumption. The distribution in existing denoising methods fail to capture an important molecular characteristic. To be specific, for 3D molecular modeling, it is essential to notice that molecules have rigid parts and flexible parts. The structures of rigid parts, such as rings, double and triple bonds, are almost fixed, whereas some single bonds can flexibly rotate. In other words, little coordinate perturbation on the rigid parts brings about high energy conformation, while altering the dihedral angles of the rotatable bonds does not cause sharp energy change. For convenience, the rule is concluded as the anisotropy of molecules or chemical constraints (Stärk et al., 2022b). Now we introduce this important chemical knowledge into the denoising framework.

3.1.1. ENLARGING SAMPLING COVERAGE

As is briefly discussed in section 1, unless the noise scale is very small, the isotropic coordinate noise will generate structures that violate the chemical constraints and thus are ineffective for downstream tasks. However, small noise scale hinders sampling more structures with low energy. Meanwhile, it is hard to mathematically define a kind of noise that both satisfies the chemical constraints and can be
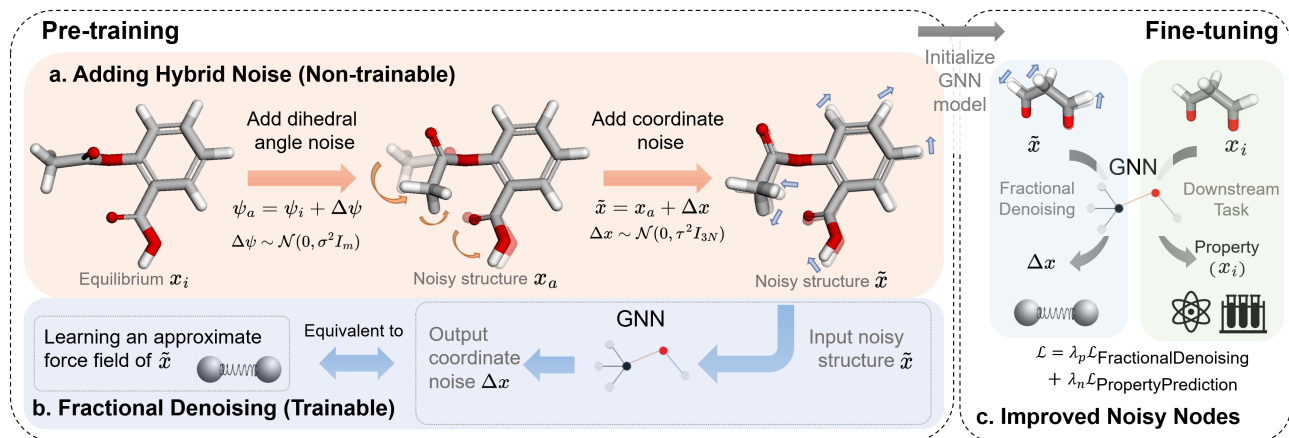
Figure 2. An overview of our method Frad. **a**: During pre-training, the hybrid noise, combining dihedral angle noise and coordinate noise, is applied to the equilibrium conformation. **b**: The GNN is trained to predict the coordinate noise, which is a fraction of the hybrid noise. This process is named Frad (Fractional Denoising), and proved to be equivalent to learning an approximate force field. **c**: We apply Frad during fine-tuning on the MD17 dataset. Specifically, fractional denoising is added as an auxiliary task, which is optimized with the primary property prediction task simultaneously.

easily implemented into the denoising framework.

Inspired by a technique utilized in molecular docking (Meng et al., 2011; Stärk et al., 2022b) and generation community (Wang et al., 2022a; Jing et al., 2022) that searches low-energy structures by generating the dihedral angles of the rotatable bonds, we propose dihedral angle noise that naturally obeys chemical constraints. In chemistry, a dihedral angle refers to the angle formed by the intersection of two half-planes through two sets of three atoms, where two of the atoms are shared between the two sets. Specifically, we search all the rotatable single bonds in the molecule and perturb the corresponding dihedral angles by Gaussian noise. These can be efficiently completed using RDKit, which is a fast cheminformatics tool (Landrum et al., 2013; Riniker & Landrum, 2015). Therefore, we can adjust the noise scale without generating invalid structures, enlarging the sampling coverage. In addition, since the diheral angle noise can generate structures with low energy, adding diheral angle noise can also be viewed as data augmentation.

3.1.2. APPROXIMATING ANISOTROPIC FORCE FIELD

Since perturbing different parts of the molecule structure can cause varying scales of effect on energy, we assume the conformation distribution $p(\tilde{x})$ has an anisotropic covariance, i.e. the slope of the energy function is not the same in all directions around the local minimum. Intuitively, the energy function should be sharp in the direction of perturbing the positions of atoms in the rigid parts, while smooth in that of the flexible parts. Correspondingly, the covariance of noise should be small on rigid parts and large on flexible parts.

Following the method in section 3.1.1, we perturb the flexible parts by turning the dihedral angles of the rotatable bonds and then perturb the whole molecule by a small level of coordinate noise, resulting in hybrid noise, as is shown in Figure 2a. In this way, the covariance of the hybrid noise is larger for the flexible parts and smaller for the rigid parts, leading to an anisotropic conformation distribution and correspondingly an anisotropic energy function that meets the chemical constraints. Therefore, the approximate conformation distribution of hybrid noise is more accurate than that of traditional coordinate noise, especially after carefully tuning the noise scales. Consequently, the approximate force field corresponding to the hybrid noise is more accurate than that of traditional coordinate noise. This is supported by our experiment in Appendix B.2. Moreover, since the noise scale of the coordinate noise is kept small, the hybrid noise still maintains the sample validity mentioned in section 3.1.1.

3.2. Difficulties to Learn the Force Field of Hybrid Noise

A challenge to the hybrid noise is that directly denoising is not equivalent to learning a force field. To better understand this difficulty, we provide the coordinate form of dihedral angle noise and hybrid noise under certain conditions.

Before providing theoretical results, we first clarify some notations. x_i denotes the equilibrium conformations, x_a denotes the conformation after perturbed by dihedral angle noise, $\tilde{x}$ denotes the conformation after perturbed by hybrid noise, $x_i, x_a, \tilde{x} \in \mathbb{R}^{3N}$, and N is the number of atoms in this molecule. If there are m rotatable bonds in the molecule, the dihedral angles of the rotatable bonds are represented by $\psi = (\psi_1, ..., \psi_m) \in [0, 2\pi)^m$. Denote $\psi_i, \psi_a, \tilde{\psi}$ as the dihedral angles of $x_i, x_a, \tilde{x}$ respectively. The notations are

consistent in Figure 2.

Proposition 3.1 (Noise Type Transformation). *Consider adding dihedral angle noise $\Delta\psi \in [0,2\pi)^m$ on the input structure x_i. The corresponding coordinate change $\Delta x = x_a - x_i \in \mathbb{R}^{3N}$ is approximately linear with respect to the dihedral angle noise, when the scale of the dihedral angle noise is small.*

$$\|\Delta x - C\Delta\psi\|_2^2 \le \sum_{j=1}^m D_j \mathcal{E}(\Delta\psi_j) \quad (3)$$

where C is a $3N \times m$ matrix that is dependent on the input conformation, $\{D_j, j = 1\cdots m\}$ are constants dependent on the input conformation. $\lim_{\Delta\psi_j \to 0} \mathcal{E}(\Delta\psi_j) = 0, \forall j = 1\cdots m$, indicating the linear approximation error is small when the scale of the dihedral angle noise is small.

The proposition in full form and its proof are in Appendix A. Proposition 3.1 provides the approximate linear relationship between the two types of noise. When the scale of the dihedral angle noise is sufficient small, Δx is sufficiently close to $C\Delta\psi$, then we get the approximate conformation distribution of x_a and $\tilde{x}$.

Proposition 3.2 (The Conformation Distribution Corresponding to Dihedral Angle Noise). *If $p(\psi_a|\psi_i) \sim \mathcal{N}(\psi_i, \sigma^2 I_m)$, i.e. Gaussian dihedral angle noise is added on the equilibrium conformation, then the approximate conformation distribution of the noisy structure x_a conditioned on equilibrium structure x_i is $p(x_a|x_i) \sim \mathcal{N}(x_i, \Sigma_\sigma)$, where $\Sigma = \sigma^2 CC^T$.*

Proposition 3.3 (The Conformation Distribution Corresponding to Hybrid Noise). *If $p(\psi_a|\psi_i) \sim \mathcal{N}(\psi_i, \sigma^2 I_m)$, $p(\tilde{x}|x_a) \sim \mathcal{N}(x_a, \tau^2 I_{3N})$ i.e. the hybrid noise is added on the equilibrium conformation, then the approximate conformation distribution of the noisy structure $\tilde{x}$ conditioned on equilibrium structure x_i is $p(\tilde{x}|x_i) \sim \mathcal{N}(x_i, \Sigma_{\sigma,\tau})$, where $\Sigma_{\sigma,\tau} = \tau^2 I_{3N} + \sigma^2 CC^T$.*

We summarize the conditional conformation distribution and corresponding conditional score function under different noise type in Table 1. Compared to traditional coordinate noise, the covariance of the hybrid noise is indeed anisotropic. In addition, the covariances in Proposition 3.2 and 3.3 are dependent on the input equilibrium structure x_i. Substitute them into the third row of (2), we have $min_\theta E_{p(\tilde{x})}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}}E(\tilde{x})\|^2 \simeq min_\theta E_{p(\tilde{x},x_i)}\|GNN_\theta(\tilde{x}) - \Sigma^{-1}(x_i - \tilde{x}))\|^2$. However, $min_\theta E_{p(\tilde{x},x_i)}\|GNN_\theta(\tilde{x}) - \Sigma^{-1}(x_i - \tilde{x}))\|^2 \neq min_\theta E_{p(\tilde{x},x_i)}\|GNN_\theta(\tilde{x}) - (x_i - \tilde{x}))\|^2$ for $\Sigma = \Sigma_\sigma$ and $\Sigma = \Sigma_{\sigma,\tau}$, i.e. neither denoising dihedral angle noise nor denoising hybrid noise is equivalent to learning the force field, because the coefficients Σ rely on input conformation and cannot be absorbed into GNN_θ. Although the theories require noise scale to be sufficiently small, it is enough to show

the difficulty to learn the force field of hybrid noise, because the equivalence should hold in all noise scale settings.

Table 1. The conditional distributions of the conformation perturbed by various noise type given the clean (equilibrium) conformation x_i and the corresponding score functions.

Noise Type	Conformation Distribution $p(\cdot|x_i)$	Score Function $\nabla \log p(\cdot|x_i)$
Coordinate	$x_{cd} \sim \mathcal{N}(x_i, \tau^2 I_{3N})$	$\frac{1}{\tau^2}(x_i - x_{cd})$
Dihedral Angle	$x_a \sim \mathcal{N}(x_i, \Sigma_\sigma)$	$\Sigma_\sigma^{-1}(x_i - x_a)$
Hybrid	$\tilde{x} \sim \mathcal{N}(x_i, \Sigma_{\sigma,\tau})$	$\Sigma_{\sigma,\tau}^{-1}(x_i - \tilde{x})$

3.3. Fractional Denoising Method

From the discussion above, we conclude that the last equation in (2) do not hold because the covariance is no longer isotropic and depends on the input structure. Note that the problem lies in the dihedral part of the hybrid noise. In order to decouple the two types of noise, we design a clever denoising method, namely fractional denoising, that adding hybrid noise while reconstructiong the coordinate part of noise, as is illustrated in Figure 2b. An exciting result is that the fractional denoising task is equivalent to learning the anisotropic force field of hybrid noise. The result is summarized below. The proof is in Proposition A.7 in appendix.

Proposition 3.4 (Fractional Denoising Score Matching). *If $p(\tilde{x}|x_a) \sim \mathcal{N}(x_a, \tau^2 I_{3N})$ and $p(x_a|x_i)$ can be arbitrary distribution, we have*

$$E_{p(\tilde{x}|x_a)p(x_a|x_i)p(x_i)}\|GNN_\theta(\tilde{x}) - (\tilde{x} - x_a)\|^2$$
$$\simeq E_{p(\tilde{x})}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}}\log p(\tilde{x})\|^2, \quad (4)$$

$\simeq$ *denotes the equivalence as optimization objectives. $\nabla_{\tilde{x}}\log p(\tilde{x}) = -\nabla_{\tilde{x}}E(\tilde{x})$ is the anisotropic force field of the hybrid noise, because $p(\tilde{x}) = \sum_{i=1}^n p(\tilde{x}|x_i)p_0(x_i)$ and $p(\tilde{x}|x_i)$ is given by hybrid noise and with an anisotropic covariance.*

Proposition 3.4 indicates the fractional denoising objective is equivalent to learning an anisotropic force field. Additionally, though we only denoise the coordinate part, Frad does not suffer from the sampling challenge because the samples $\tilde{x}$ are generated by hybrid noise. Besides, $p(x_a|x_i)$ can be arbitrary distribution, leaving room for designing more accurate energy functions in future work. In particular, fractional denoising is implemented as follows: our model takes $\tilde{x}$ as input and predicts $(\tilde{x} - x_a)$ as the denoising target. We train our network GNN_θ to minimize the pre-training loss function defined in Equation 5. For a complete description of our method's pipeline, please refer to Algorithm 1 in

Table 2. Performance (MAE, lower is better) on QM9 force prediction. The best results are in bold.

Models	μ (D)	α (a_0^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	$\Delta\epsilon$ (meV)	$<R^2>$ (a_0^2)	ZPVE (meV)	U_0 (meV)	U (meV)	H (meV)	G (meV)	C_v $(\frac{cal}{molK})$
SchNet	0.033	0.235	41.0	34.0	63.0	0.07	1.70	14.00	19.00	14.00	14.00	0.033
E(n)-GNN	0.029	0.071	29.0	25.0	48.0	0.11	1.55	11.00	12.00	12.00	12.00	0.031
DimeNet++	0.030	0.043	24.6	19.5	32.6	0.33	1.21	6.32	6.28	6.53	7.56	0.023
PaiNN	0.012	0.045	27.6	20.4	45.7	0.07	1.28	5.85	5.83	5.98	7.35	0.024
SphereNet	0.027	0.047	23.6	18.9	32.3	0.29	**1.120**	6.26	7.33	6.40	8.00	0.022
TorchMD-NET	0.011	0.059	20.3	18.6	36.1	**0.033**	1.840	6.15	6.38	6.16	7.62	0.026
Transformer-M	0.037	0.041	17.5	16.2	**27.4**	0.075	1.18	9.37	9.41	9.39	9.63	0.022
SE(3)-DDM	0.015	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024
3D-EMGP	0.020	0.057	21.3	18.2	37.1	0.092	1.38	8.60	8.60	8.70	9.30	0.026
DP-TorchMD -NET($\tau = 0.04$)	0.012	0.0517	17.7	14.3	31.8	0.4496	1.71	6.57	6.11	6.45	6.91	**0.020**
Frad ($\sigma = 2, \tau = 0.04$)	**0.010**	**0.0374**	**15.3**	**13.7**	27.8	0.3419	1.418	**5.33**	**5.62**	**5.55**	**6.19**	**0.020**

appendix.

$$\mathcal{L}_{pre-training} =$$
$$E_{p(\tilde{x}|x_a)p(x_a|x_i)p(x_i)}\|GNN_\theta(\tilde{x}) - (\tilde{x} - x_a)\|_2^2. \quad (5)$$

3.4. Applying Frad to Fine-tuning

In addition to pre-training, (Godwin et al., 2021; Zaidi et al., 2022) show that denoising can also improve representation learning in fine-tuning by discouraging over-smoothing and learning data distribution. The method is called Noisy Nodes, which incorporates an auxiliary loss for coordinate denoising in addition to the original property prediction objective. Specifically, it corrupts the input structure by coordinate noise and then trains the model to predict the properties and the noise from the same noisy structure. We have included pseudocode for Noisy Nodes in Algorithm 2, provided for reference. Unfortunately, we find it cannot converge on tasks in MD17 dataset. We conjecture that this is because the task in MD17 is sensitive to the input conformation (see section 4.1.1), whereas Noisy Nodes have to corrupt the input conformation leading to an erroneous mapping between inputs and property labels.

To fill this gap, we propose two modifications, as is illustrated in Figure 2c. For one thing, we decouple the denoising task and the downstream task to keep the input of the downstream task unperturbed. The two tasks are trained simultaneously by optimising a weighted sum of losses of the two tasks. For the other thing, we substitute our hybrid noise and fractional denoising for the coordinate denoising in Noisy Nodes, so the benefit of force field learning can also be inherited. Specifically, Equation 6 defines the optimization goal of our modified Noisy Nodes.

$$\mathcal{L}_{fine-tuning} = \lambda_p \mathcal{L}_{FractionalDenoising}$$
$$+ \lambda_n \mathcal{L}_{PropertyPrediction} \quad (6)$$

where $\mathcal{L}_{FractionalDenoising} = E_{p(\tilde{x}|x_a)p(x_a|x_i)p(x_i)}\|\Delta x_i^{pred} - (\tilde{x} - x_a)\|_2^2$, $\Delta x_i^{pred} = \text{NoiseHead}_{\theta_n}(GNN_\theta(\tilde{x}))$ denotes the prediction of noise, $\mathcal{L}_{PropertyPrediction} = PropertyPredictionLoss(y_i^{pred}, y_i)$ that can be in different form for various downstream tasks, $y_i^{pred} = \text{LabelHead}_{\theta_l}(GNN_\theta(x_i))$ represents the predicted label of x_i, λ_p and λ_n represent the loss weight of property prediction and Noisy Nodes respectively, the $\text{NoiseHead}_{\theta_n}$ module takes the representation of $\tilde{x}$ as its input and generates a predicted node-level noise for each atom, while the $\text{LabelHead}_{\theta_l}$ employs the representation of x_i to forecast the graph-level label for x_i. The full optimization pipeline can be found in Algorithm 3. Ablation study in section 4.3.2 validates that both of the modifications contribute to better performance. Our modified Noisy Nodes may further benefit the tasks that are sensitive to the input conformations, such as ligand generation, affinity prediction, and so on. We leave it as future work.

4. Experiments

4.1. Settings

4.1.1. DATASETS

We leverage a large-scale molecular dataset PCQM4Mv2 (Nakata & Shimazaki, 2017) as our pre-training dataset. It contains 3.4 million organic molecules, with one equilibrium conformation and one label calculated by density functional theory (DFT). We do not use the label since our method is self-supervised.

As for downstream tasks, we adopt the two popular 3D molecular property prediction datasets: MD17 (Chmiela et al., 2017) and QM9 (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014). QM9 is a quantum chemistry dataset including geometric, energetic, electronic and thermodynamic properties for 134k stable small organic molecules

Table 3. Performance (MAE, lower is better) on MD17 force prediction. The best results are in bold. *The result is reported by SE(3)-DDM and PaiNN does not provide the result on Benzene.

Training set size	Models	Aspirin	Benzene	Ethanol	Malonaldehyde	Naphthalene	Salicylic Acid	Toluene	Uracil
9500	TorchMD-NET	0.1216	0.1479	0.0492	0.0695	0.0390	0.0655	0.0393	0.0484
	3D-EMGP	0.1560	0.1648	0.0389	0.0737	0.0829	0.1187	0.0619	0.0773
	3D-EMGP (TorchMD-NET)	0.1124	0.1417	0.0445	0.0618	0.0352	0.0586	0.0385	0.0477
	DP-TorchMD -NET($\tau = 0.04$)	0.0920	0.1397	0.0402	0.0661	0.0544	0.0790	0.0495	0.0507
	Frad ($\sigma = 2, \tau = 0.04$)	**0.0680**	0.1606	**0.0332**	**0.0427**	**0.0277**	**0.0410**	**0.0305**	**0.0323**
1000	SphereNet	0.430	0.178	0.208	0.340	0.178	0.360	0.155	0.267
	SchNet	1.35	0.31	0.39	0.66	0.58	0.85	0.57	0.56
	DimeNet	0.499	0.187	0.230	0.383	0.215	0.374	0.216	0.301
	SE(3)-DDM	0.453	**0.051**	0.166	0.288	0.129	0.266	0.122	0.183
950	PaiNN	0.338	0.052*	0.224	0.319	0.077	0.195	0.094	0.139
	TorchMD-NET	0.2450	0.2187	0.1067	0.1667	0.0593	0.1284	0.0644	0.0887
	Frad ($\sigma = 2, \tau = 0.04$)	**0.2087**	0.1994	**0.0910**	**0.1415**	**0.0530**	**0.1081**	**0.0540**	**0.0760**

made up of CHONF atoms. Each molecule has one equilibrium conformation and 12 labels calculated by density functional theory (DFT). The QM9 dataset is split into a training set with 110,000 and a validation set with 10,000 samples, leaving 10,831 samples for testing. This splitting is commonly applied in literature. As usually done on QM9, we fine-tune a separate model for each of the 12 downstream tasks, with the same pre-trained model.

MD17 is a dataset of molecular dynamics trajectories, containing 8 small organic molecules with conformations, total energy and force labels computed by electronic structure method. For each molecule, 150k to nearly 1M conformations are provided. Therefore, compared to QM9, the property prediction task of MD17 is more sensitive to the input conformations. We note that the force prediction task is more discriminative and widely-used than the energy prediction task. So we choose force prediction as the downstream task. Regarding data splitting, the approaches diverge on taking large (9500) or small (950 or 1000) size of training data. As the size of training dataset affects the force prediction significantly, we perform Frad with both splitting for fair comparisons. More settings are summarized in Appendix B.5.

4.1.2. BASELINES

In terms of 3D pre-training approaches, our baselines cover the currently SOTA methods we have known, including DP-TorchMD-NET (Zaidi et al., 2022), 3D-EMGP (Jiao et al., 2022), SE(3)-DDM (Liu et al., 2022a), Transformer-M (Luo et al., 2022). DP-TorchMD-NET is the baseline we are most interested in, because it is a typical coordinate denoising pre-training method and shares the same backbone with Frad.

So their performance well reflects the comparison between coordinate denoising and fractional denoising. As equivariant denoising methods, 3D-EMGP and SE(3)-DDM are also important baselines to judge whether the prior knowledge we incorporate and the way we incorporate are better. As for Transformer-M, it is a competitive model consisting of denoising and energy prediction pre-training tasks. We exclude Uni-mol (Zhou et al., 2022) and ChemRL-GEM (Fang et al., 2021) since they only provide the average performance of 3 energy tasks in QM9.

We also adopt the representative approaches designed for property prediction to test our ability as a property prediction model. The approaches are not pre-trained and they comprise TorchMD-NET (Thölke & De Fabritiis, 2022), SchNet (Schütt et al., 2018), E(n)-GNN(Satorras et al., 2021), DimeNet (Gasteiger et al., 2020), DimeNet++ (Klicpera et al., 2020), SphereNet (Liu et al., 2022b), PaiNN (Schütt et al., 2021). Among them, we employ TorchMD-NET as our backbone, which is an equivariant Transformer architecture for 3D inputs. For fair comparison with coordinate denoising, we use the publicly available code from Zaidi et al. (2022) to produce results for DP-TorchMD-NET. The result of TorchMD-NET with 9500 training data of MD17 is reported by (Jiao et al., 2022). Other results are taken from the referred papers.

4.2. Main Experimental Result

4.2.1. RESULTS ON QM9

In this section, we evaluate the models on QM9 and verify whether Frad can consistently achieve competitive results. The performance is measured by mean absolute error (MAE)

for each property and the results are summarized in Table 2.

In general, we achieve a new state-of-the-art for 9 out of 12 targets. The models for the upper half of the table are property prediction baselines without pre-training. We exceed them on most of the tasks. It is worth mentioning that we make remarkable improvements on the basis of the backbone TorchMD-NET on 11 targets, indicating the effectiveness of our method. As for the outlier $< R^2 >$, we observe the same phenomenon in DP-TorchMD-NET. We speculate it is because the optimal noise scale of $< R^2 >$ is different from that of other targets.

We also have an evident advantage over the denoising pre-training methods in the lower half of the table. Especially, our Frad achieves or surpasses the results of coordinate denoising approach DP-TorchMD-NET in all 12 tasks, revealing that chemical constraints are unneglectable in denoising. Here DP-TorchMD-NET is trained with hyperparameters in the code of Zaidi et al. (2022). A comparison with strictly aligned setting between coordinate denoising and Frad is in section 4.3.1.

4.2.2. RESULTS ON MD17

Compared with QM9, tasks in MD17 are more sensitive to molecular geometry and contain nonequilibrium conformations, which bring new challenges to the models. Theoretically, denoising can directly benefit downstream force learning, since it has learnt an approximate force field as a reference. As we expected, Frad achieves new SOTA and the results are in Table 3. In both large and small training data scenarios, Frad outperforms the corresponding pretrained and non-pretrained baselines on 7 out of 8 molecules. Especially when comparing with 3D-EMGP(TorchMD-NET) and DP-TorchMD-NET who utilize the same backbone as us, our superiority is evident, showing the necessity to correct denoising methods by chemical constraints.

Regarding Benzene, we observe overfitting during fine-tuning the Frad, which is not found in other molecules. This may be caused by the relatively fixed structure of benzene, leading to low-dimensional features which are easy to overfit. In addition, we can see from the table that the best result on Benzene, achieved by SE(3)-DDM, may be mainly attributed to the backbone PaiNN. Correspondingly, the inferior performance of Frad may come from the backbone TorchMD-NET rather than denoising.

4.3. Ablation Study

The Frad technique can be applied in pre-training phase as training targets and fine-tuning phase by Noisy Nodes. Then how much does each part contribute to the final result? Our ablation study here validates each part respectively.

4.3.1. FRAD IN PRE-TRAINING

To verify Frad as effective pre-training target, we evaluate Frad and coordinate denoising on 6 tasks in QM9. The settings for the two approaches, including hyperparameters for optimization, network structure and Noisy Nodes, are strictly aligned in each task. The results are displayed in Table 4. Frad surpasses coordinate denoising on all 6 tasks, indicating the significance of chemical constraints in force field learning. Note that QM9 contains multiple categories of equilibrium properties, including thermodynamic properties, spatial distribution of electrons and states of the electrons. We speculate an accurate force field learning can not only assist energy prediction, but may enhance the atomic charge prediction and its related properties as well.

Table 4. The performance (MAE) of coordinate denoising and Frad on QM9. The top results are in bold.

	μ (D)	α (a_0^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	$\Delta\epsilon$ (meV)	$< R^2 >$ (a_0^2)
$\tau = 0.04$	0.0120	0.0517	17.7	14.3	31.8	0.4496
$\sigma = 2$, $\tau = 0.04$	**0.0118**	**0.0486**	**15.3**	**13.7**	**27.8**	**0.4265**

4.3.2. FRAD IN FINE-TUNING

Next, to validate our improvements on Noisy Nodes, we use the same model pre-trained by Frad ($\sigma = 2$, $\tau = 0.04$) and fine-tune it on Aspirin task in MD17 with distinct Noisy Nodes settings. The results are in Table 5. The analysis are threefold. Setting 2-6 v.s. setting 1: The traditional Noisy Nodes fail to converge while our modifications fix the problem. Setting 3-4 v.s. setting 2: Setting 3 can converge because the dihedral angle noise has less influence on the energy. Additionally, decoupling the input of different tasks ensures an unperturbed input for property prediction, and fundamentally corrects the mapping, allowing setting 4 to work effectively. Setting 5 v.s. setting 4: Fractional denoising further promotes the performance of Noisy Nodes. Combined with experiments in section B.2, we can infer that learning a more accurate force field indeed contributes to downstream tasks.

5. Conclusion

This paper is concerned with coordinate denoising approach for 3D molecule pretraining. We find that existing coordinate denoising methods has two major limitations, i.e. low sampling coverage and isotropic force field, which prevent the current methods from learning an accurate force filed. To tackle these challenges, we propose a novel denoising method, namely Frad. By introducing hybrid noises on both

Table 5. Performance (MAE, lower is better) of different fine-tuning techniques on Aspirin task in MD17. NN denotes Noisy Nodes. DEC stands for decoupling the input of denoising and downstream tasks. The best results are in bold.

Index	Settings	Aspirin (Force)
1	w/o NN	0.2141
2	NN($\tau = 0.005$)	do not converge
3	NN($\sigma = 0.2$)	0.2096
4	NN($\tau = 0.005$, DEC)	0.2107
5	NN($\sigma = 20$, $\tau = 0.005$, DEC)	**0.2087**

dihedral angel and coordinate, Frad has the ability to sample more low-energy conformations. Besides, by denoising only the coordinate noise, Frad is proven to be equivalent to a more reasonable anisotropic force field. Consequently, Frad achieves new SOTA on QM9 and MD17 as compared with existing coordinate denoising methods. Ablation studies show the superiority of Frad over coordinate denoising in terms of both pre-training and fine-tuning.

Our work has provided several potential directions. Firstly, Proposition 3.4 holds without limiting the angle noise type, suggesting fractional denoising could be a general technique worth indepth investigation. Secondly, here is another point of view to understand our Frad that the dihedral angle noise is a data augmentation strategy to search more low energy structures, while the fractional denoising method is purposed to learning an effective molecular representation insensitive to the coordinate noise. This perspective may inspire new pre-training methods based on both contrastive learning and denoising. Thirdly, how to design a denoising method which better captures the characteristics of molecules to learn a more accurate force field is still an open question.

Acknowledgements

References

Asada, M., Miwa, M., and Sasaki, Y. Enhancing drug-drug interaction extraction from texts by molecular structure information. *arXiv preprint arXiv:1805.05593*, 2018.

Bishop, C. M. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 1995.

Boltzmann, L. Studien uber das gleichgewicht der lebenden kraft. *Wissenschafiliche Abhandlungen*, 1:49–96, 1868.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

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

Chmiela, S., Sauceda, H. E., Poltavsky, I., Müller, K.-R., and Tkatchenko, A. sgdml: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications*, 240:38–45, 2019.

Dai, A. M. and Le, Q. V. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28, 2015.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Chemrl-gem: Geometry enhanced molecular representation learning for property prediction. *arXiv preprint arXiv:2106.06130*, 2021.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022a.

Fang, Y., Zhang, Q., Yang, H., Zhuang, X., Deng, S., Zhang, W., Qin, M., Chen, Z., Fan, X., and Chen, H. Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3968–3976, 2022b.

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

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

Godwin, J., Schaarschmidt, M., Gaunt, A., Sanchez-Gonzalez, A., Rubanova, Y., Veličković, P., Kirkpatrick, J., and Battaglia, P. Simple gnn regularisation for 3D molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971*, 2021.

Guo, Z., Sharma, P., Martinez, A., Du, L., and Abraham, R. Multilingual molecular representation learning via contrastive pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3441–3453, 2022.

Honda, S., Shi, S., and Ueda, H. R. Smiles transformer: Pretrained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3D. *international conference on machine learning*, 2023.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Jiao, R., Han, J., Huang, W., Rong, Y., and Liu, Y. Energy-motivated equivariant pretraining for 3D molecular graphs. *arXiv preprint arXiv:2207.08824*, 2022.

Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

Klicpera, J., Giri, S., Margraf, J. T., and Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020.

Kong, K., Li, G., Ding, M., Wu, Z., Zhu, C., Ghanem, B., Taylor, G., and Goldstein, T. Flag: Adversarial data augmentation for graph neural networks. *arXiv preprint arXiv:2010.09891*, 2020.

Landrum, G. et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 2013.

Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., Gao, P., Xie, G., and Song, S. Learn molecular representations from large-scale unlabeled molecules for drug discovery. *arXiv preprint arXiv:2012.11175*, 2020.

Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., Gao, P., Xie, G., and Song, S. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings in Bioinformatics*, 22 (6):bbab109, 2021.

Li, S., Zhou, J., Xu, T., Dou, D., and Xiong, H. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4541–4549, 2022.

Lin, X., Xu, C., Xiong, Z., Zhang, X., Ni, N., Ni, B., Chang, J., Pan, R., Wang, Z., Yu, F., et al. Pangu drug model: Learn a molecule like a human. *bioRxiv*, 2022.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3D geometry. *arXiv preprint arXiv:2110.07728*, 2021.

Liu, S., Guo, H., and Tang, J. Molecular geometry pretraining with SE(3)-invariant denoising distance matching. 2022a.

Liu, Y., Wang, L., Liu, M., Lin, Y., Zhang, X., Oztekin, B., and Ji, S. Spherical message passing for 3D molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2022b.

Luo, S., Chen, T., Xu, Y., Zheng, S., Liu, T.-Y., Wang, L., and He, D. One transformer can understand both 2D & 3D molecular data. *arXiv preprint arXiv:2210.01765*, 2022.

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

Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7 (2):146–157, 2011.

Nakata, M. and Shimazaki, T. Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. *Journal of chemical information and modeling*, 57(6):1300–1308, 2017.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Riniker, S. and Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling*, 55(12):2562–2574, 2015.

Rohani, N. and Eslahchi, C. Drug-drug interaction predicting by neural network using integrated similarity. *Scientific reports*, 9(1):1–11, 2019.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Sato, R., Yamada, M., and Kashima, H. Random features strengthen graph neural networks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 333–341. SIAM, 2021.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Schütt, K., Unke, O., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pp. 9377–9388. PMLR, 2021.

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

Sietsma, J. and Dow, R. J. Creating artificial neural networks that generalize. *Neural networks*, 4(1):67–79, 1991.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *neural information processing systems*, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., and Liò, P. 3D infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pp. 20479–20502. PMLR, 2022a.

Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., and Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pp. 20503–20521. PMLR, 2022b.

Thölke, P. and De Fabritiis, G. Torchmd-net: Equivariant transformers for neural network based molecular potentials. *arXiv preprint arXiv:2202.02541*, 2022.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 2011.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. *international conference on machine learning*, 2008.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 2010.

Wang, L., Zhou, Y., Wang, Y., Zheng, X., Huang, X., and Zhou, H. Regularized molecular conformation fields. In *Advances in Neural Information Processing Systems*, 2022a.

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. Smilesbert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pp. 429–436, 2019.

Wang, Y., Magar, R., Liang, C., and Barati Farimani, A. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *Journal of Chemical Information and Modeling*, 2022b.

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022c.

Xue, D., Zhang, H., Xiao, D., Gong, Y., Chuai, G., Sun, Y., Tian, H., Wu, H., Li, Y., and Liu, Q. X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis. *bioRxiv*, pp. 2020–12, 2021.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. 2022.

Zhang, X.-C., Wu, C.-K., Yang, Z.-J., Wu, Z.-X., Yi, J.-C., Hsieh, C.-Y., Hou, T.-J., and Cao, D.-S. Mg-bert: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings in bioinformatics*, 22(6):bbab152, 2021a.

Zhang, Z., Liu, Q., Wang, H., Lu, C., and Lee, C.-K. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021b.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: A universal 3D molecular representation learning framework. 2022.

Zhu, J., Xia, Y., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. Dual-view molecule pre-training. *arXiv preprint arXiv:2106.10234*, 2021.

Zhu, J., Xia, Y., Wu, L., Xie, S., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. Unified 2D and 3D pre-training of molecular representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2626–2636, 2022.

A. Proofs of Propositions

Proposition A.1 (Noise type transformation). *Consider adding dihedral angle noise $\Delta\psi$ on the input structure x_i. The corresponding coordinate change $\Delta x = x_a - x_i$ is approximately linear with respect to the dihedral angle noise, when the scale of the dihedral angle noise is small.*

$$\|\Delta x - C\Delta\psi\|_2^2 \le \sum_{j=1}^m D_j \mathcal{E}(\Delta\psi_j) \tag{7}$$

where $\Delta\psi \in [0, 2\pi)^m$, $\Delta x \in \mathbb{R}^{3N}$, C is a $3N \times m$ matrix that is dependent on the input conformation, $\{D_j, j = 1 \cdots m\}$ are constants dependent on the input conformation. $\lim_{\Delta\psi_j \to 0} \mathcal{E}(\Delta\psi_j) = \lim_{\Delta\psi_j \to 0}[\Delta\psi_j^2 - 2\Delta\psi_j \sin\Delta\psi_j - 2cos\Delta\psi_j + 2] = 0, \forall j = 1\cdots m$, indicating the linear approximation error is small when the scale of the dihedral angle noise is small. If we further assume the rotations are not trivial, i.e. rotation of the rotatable bonds causes the movement of some atom positions, the rank of C is m. All the elements above the main diagonal in C are zero.

Proof. To analyze the coordinate change after altering the dihedral angles of all the rotatable bonds in a molecule, we first consider changing one dihedral angle. As a proof in elementary geometry, we define some notations: $\overline{AA'}$ is the length of the line segment AA', $\widehat{AA'}$ is the length of the arc segment AA', $\angle$ represents angle, $\triangle A'AA''$ represents a triangle formed by point A', A and A",$\perp$ represents perpendicular.

Lemma A.2. *If the change in one dihedral angle $\Delta\psi$ is small enough, the distance the associated atoms move is proportional to the amount of the change in that dihedral angle, and the proportional coefficient is dependent on the input conformation.*

Proof. For example, in Figure 3, we study the effect on the coordinates of atom A of turning the dihedral angle ψ_1, i.e. $\angle AO_A A' = \Delta\psi_1$. If the scale of the dihedral angle noise $\Delta\psi_1$ is small, the distance can be approximated by arc length.

$$\overline{AA'} \approx \widehat{AA'} = (\overline{OA}\cos\angle OAO_A)\Delta\psi_1, \tag{8}$$

Please note that $\overline{OA}$ and $\angle OAO_A$ are both determined by the original structure and remain constant when changing the dihedral angle. Therefore, $\overline{AA'} \propto \Delta\psi_1$.

The same can be proved for other associated atoms. For example for atom B, with $\angle BO_B B' = \Delta\psi_1$ and $\overline{BB'} \approx \widehat{BB'} = (\overline{OB}\cos\angle OBO_B)\Delta\psi_1$, we can deduce $\overline{BB'} \propto \Delta\psi_1$.

In conclusion, the distance the associated atoms move is proportional to the amount of change in the dihedral angle. $\square$

Then, we extend this conclusion from distance to coordinates. Note that in lemma A.3, the notation x represents one coordinate of the 3D coordinates, relative to y and z.

Lemma A.3. *If the change in one dihedral angle $\Delta\psi$ is small enough so that the distance the associated atoms move is also small, then the coordinate changes of the associated atoms are also proportional to the amount of the change in that dihedral angle, and the proportional coefficient is dependent on the input conformation.*

Proof. Denote the 3D coordinate of atom A as (x, y, z). When changing the dihedral angle, the atom lies on the circle with center O_A, radius $\overline{O_A A}$ and perpendicular to $\overline{O_A O}$, i.e. (x, y, z) satisfies

$$\begin{cases} (x - x_{O_A})^2 + (y - y_{O_A})^2 + (z - z_{O_A})^2 = \overline{O_A A}^2 & (9) \\ (x - x_{O_A})(x_O - x_{O_A}) + (y - y_{O_A})(y_O - y_{O_A}) + (z - z_{O_A})(z_O - z_{O_A}) = 0. & (10) \end{cases}$$

Considering a sufficient small amount of coordinate movement, (x, y, z) also satisfies the formula after differentiating the equation (9)(10).

$$\begin{cases} 2(x - x_{O_A})\Delta x + 2(y - y_{O_A})\Delta y + 2(z - z_{O_A})\Delta z = 0 & (11) \\ (x_O - x_{O_A})\Delta x + (y_O - y_{O_A})\Delta y + (z_O - z_{O_A})\Delta z = 0. & (12) \end{cases}$$
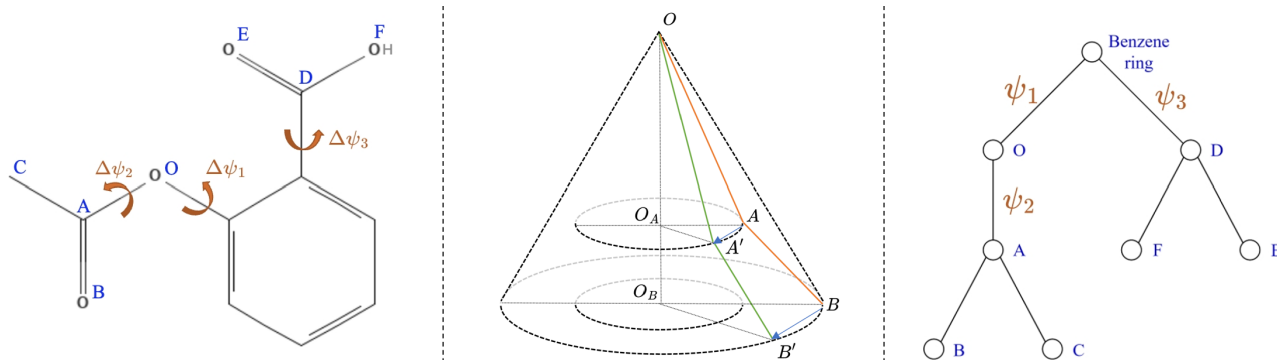
13

Figure 3. Illustrations to aid the proof of Proposition A.1. Left: Three rotatable bonds in aspirin. Middle: When changing the dihedral angle ψ_1, the atoms move along a circular arc e.g. $A \to A'$ and $B \to B'$. Right: When considering the changing of all dihedral angles, we can define a breadth-first order to traverse all rotatable bonds in the tree structure of aspirin, e.g. (ψ_1, ψ_3, ψ_2), and consider their effects on coordinate one by one and then add the effects together.

Since $OA \perp OO_A$, (11)(12) are not linearly related. Then $\Delta x, \Delta y, \Delta z$ are in linear relationship with each other.

$$\begin{cases} \Delta y = C_x^y \Delta x & (13) \\ \Delta z = C_x^z \Delta x, & (14) \end{cases}$$

where the constants $C_x^y = -\frac{(x-x_{O_A})(z_O-z_{O_A})-(x_O-x_{O_A})(z-z_{O_A})}{(y-y_{O_A})(z_O-z_{O_A})-(y_O-y_{O_A})(z-z_{O_A})}$, $C_x^z = \frac{(x-x_{O_A})(y_O-y_{O_A})-(x_O-x_{O_A})(y-y_{O_A})}{(y-y_{O_A})(z_O-z_{O_A})-(y_O-y_{O_A})(z-z_{O_A})}$.
With lemma A.2 and the relationship between the coordinates and the distance, we obtain

$$(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 = \overline{AA'}^2 \propto \Delta\psi_1^2. \quad (15)$$
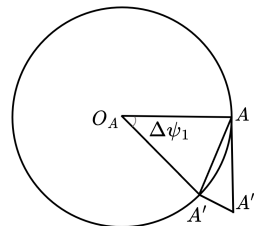
Substituting (13)(14) into (15), we conclude

$$\Delta x \propto \Delta y \propto \Delta z \propto \Delta\psi_1.$$

The same can be proved for other associated atoms. Therefore the coordinate changes are proportional to the amount of the change in the dihedral angle. $\square$

The proof in lemma A.2 and lemma A.3 requires "$\Delta\psi_1$ is sufficiently small", indicating that the linear relationship between two types of noise is an approximation. Here we specify the approximation error.

Lemma A.4. *The approximation error for atom A is given by $\overline{O_A A}^2 \mathcal{E}(\Delta\psi_1) = (\overline{OA}\cos\angle OAO_A)^2 \mathcal{E}(\Delta\psi_1)$, where $\mathcal{E}(\Delta\psi_1) = [\Delta\psi_1^2 - 2\Delta\psi_1\sin\Delta\psi_1 - 2cos\Delta\psi_1 + 2]$ is the term only dependent on $\Delta\psi_1$ in the error. $\lim_{\Delta\psi_1\to 0}\mathcal{E}(\Delta\psi_1) = 0$, indicating the approximation error is small when $\Delta\psi_1$ is small. $\overline{O_A A} = \overline{OA}\cos\angle OAO_A$ is determined by the molecular structure.*

Proof. The approximation used in (8), (11), (12) is summarized by approximating $\vec{AA'}$ by $\vec{AA''}$ in Figure 4, where $\vec{AA'}$ is the coordinate change after adding dihedral angle noise while $\vec{AA''}$ is the approximated coordinate change that is linear to the dihedral angle noise, $\overline{AA''} = \widehat{AA'}$ (i.e. (8)) and AA'' is the tangent to circle O_A at point A (i.e. (11), (12)). Therefore, $\overline{AA'} = 2\overline{O_A A}sin(\frac{\Delta\psi_1}{2})$, $\angle A'AA'' = \frac{\Delta\psi_1}{2}$,

$\overline{AA''} = \widehat{AA'} = \overline{O_A A}\Delta\psi_1$. By using the law of cosines in $\triangle A'AA''$, we have the approximating error $\|\vec{AA'} - \vec{AA''}\|_2^2 = \|\vec{A''A'}\|_2^2 = \overline{A'A''}^2 = \overline{O_A A}^2[(2sin(\frac{\Delta\psi_1}{2}))^2 + (\Delta\psi_1)^2 - 2(2sin(\frac{\Delta\psi_1}{2}))\Delta\psi_1cos(\frac{\Delta\psi_1}{2})] = \overline{O_A A}^2[2(1-cos(\Delta\psi_1)) + (\Delta\psi_1)^2 -$ *Figure 4.* Illustrations to aid the proof of $2\Delta\psi_1sin(\Delta\psi_1)]$. $\lim_{\Delta\psi_1\to 0}[2(1-cos(\Delta\psi_1))+(\Delta\psi_1)^2-2\Delta\psi_1sin(\Delta\psi_1)] = 0$. $\square$ Lemma A.4

Now we consider changing all the dihedral angles of the rotatable bonds. Making use of lemma A.3, the overall changes in coordinates are the sum of the coordinate change caused by each dihedral angle. Denote the linear coefficients as C, we can obtain (7), and C is dependent on the input conformation.

In fact, by specifying the form of $\Delta\psi$ and Δx in (7), we can further prove all the elements above the main diagonal in C are zero and full rank. One the one hand, a molecule forms a tree structure, if rings and other atoms in the molecule are regarded as nodes, and the bonds are regarded as edges. We can define a breadth-first order to traverse all rotatable bonds. We arrange the dihedral angles in the vector $\Delta\psi$ in this order. One the other hand, when rotating each bonds sequentially, we consider the effects of the dihedral angles on its child nodes and keep the parent nodes still. Define one of the nearest atoms affected by the dihedral angle as the key atom of the dihedral angle. We put the coordinates of the key atoms in the first few components in the vector Δx in (7) and arrange them correspondingly in the order of the dihedral angles they belong to. All the elements above the main diagonal in C are zero because in the breadth-first order, later dihedral angles do not affect key atoms of the earlier dihedral angles. Also, we assume the rotations are not trivial, so every rotation of the rotatable bonds has a key atom and thus the diagonal blocks (3×1 submatrices) of C are not 0 submatrices (3×1 submatrix whose elements are all zeros). Coupled with the condition that the elements above the main diagonal are zero, we derive the conclusion that the matrix C is full rank.

Take aspirin as an example, as is shown in Figure 3. Without loss of generality, we discuss three rotatable bonds labeled in the figure. Represent aspirin as a tree (the representation is not unique), and traverse all rotatable bonds in breadth-first order (ψ_1, ψ_3, ψ_2). Their corresponding key atoms are (A, F, B). For simplicity, denote $\Delta_k \triangleq (\Delta_{x_k}, \Delta_{y_k}, \Delta_{z_k})^T$ as coordinate changes of atom k, $k \in \{A, B, C, E, F\}$, and denote Δ_{Other} as coordinate changes of the atoms whose coordinates remain unchanged after altering the rotating the rotatable bonds. Then the approximate relation between atomic coordinate changes and dihedral angle changes is given by

$$
\begin{pmatrix} \Delta_A \\ \Delta_F \\ \Delta_B \\ \Delta_C \\ \Delta_E \\ \Delta_{Other} \end{pmatrix} \approx \begin{pmatrix} c_{11} & 0 & 0 \\ 0 & c_{22} & 0 \\ c_{31} & 0 & c_{33} \\ c_{41} & 0 & c_{43} \\ 0 & c_{52} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta\psi_1 \\ \Delta\psi_3 \\ \Delta\psi_2 \end{pmatrix},
\tag{16}
$$

where $c_{ij}, i, j = 1, 2\cdots$ represents 3×1 submatrices and 0 in the matrix represents 3×1 submatrix whose elements are all zeros.

The approximation error $\|\Delta x - C\Delta\psi\|_2^2 = \sum_i \|\Delta_{atom_i} - \sum_j c_{ij}\Delta\psi_j\|_2^2 = \sum_i \|\sum_j \Delta_{atom_i,\psi_j} - \sum_j c_{ij}\Delta\psi_j\|_2^2 \leq \sum_{i,j} \|\Delta_{atom_i,\psi_j} - c_{ij}\Delta\psi_j\|_2^2 = \sum_{i,j} d_{atom_i,\psi_j}^2 \mathcal{E}(\Delta\psi_j) = \sum_j (\sum_i d_{atom_i,\psi_j}^2)\mathcal{E}(\Delta\psi_j)$, where Δ_{atom_i} is the total coordinate change for $atom_i$, Δ_{atom_i,ψ_j} is the coordinate change caused by $\Delta\psi_j$ for $atom_i$ (It varies with the order of changing the dihedral angles, but the sum $\Delta_{atom_i} = \sum_j \Delta_{atom_i,\psi_j}$ is unique defined.), d_{atom_i,ψ_j} is the radius of the circular motion of atom i when the dihedral angle of ψ_j changes, e.g. $\overline{O_A A}$ in lemma A.4. $\qquad\square$

Proposition A.5. *Denote x_a to be the noisy conformation after adding Gaussian dihedral angle noise on the equilibrium conformation x_i, i.e. $p(\psi_a|\psi_i) \sim \mathcal{N}(\psi_i, \sigma^2 I_m)$, then the approximate conditional probability distribution is*

$$
p(x_a|x_i) \sim \mathcal{N}(x_i, \sigma^2 CC^T).
\tag{17}
$$

Proof.

$$
x_a = x_i + \Delta x \approx x_i + C\Delta\psi.
\tag{18}
$$

Since a linear transformation of a Gaussian random variable is still a Gaussian random variable, $p(x_a|x_i)$ is a Gaussian random variable.

$$
cov(x_a) = cov(\Delta x) \approx cov(C\Delta\psi) = Ccov(\Delta\psi)C^T = \sigma^2 CC^T,
\tag{19}
$$

therefore $p(x_a|x_i) \sim \mathcal{N}(x_i, \sigma^2 CC^T)$. $\qquad\square$

Proposition A.6. *If $p(\psi_a|\psi_i) \sim \mathcal{N}(\psi_i, \sigma^2 I_m)$, $p(\tilde{x}|x_a) \sim \mathcal{N}(x_a, \tau^2 I_{3N})$ i.e. the hybrid noise is added on the equilibrium conformation, then the approximate conformation distribution of the noisy structure $\tilde{x}$ conditioned on equilibrium structure x_i is $p(\tilde{x}|x_i) \sim \mathcal{N}(x_i, \Sigma_{\sigma,\tau})$, where $\Sigma_{\sigma,\tau} = \tau^2 I_{3N} + \sigma^2 CC^T$.*

15

Proof. We can parameterize $\tilde{x}$ to be $x_a + \tau\epsilon$, where $\epsilon \sim \mathcal{N}(0, I_{3N})$. Then $\tilde{x} \approx x_i + \tau\epsilon + C\Delta\psi$, where $\Delta\psi \sim \mathcal{N}(0, \sigma^2 I_m)$, $\epsilon \perp\!\!\!\perp \Delta\psi$. Since the sum of independent Gaussian random variables is still a Gaussian random variable, $p(\tilde{x}|x_i)$ is a Gaussian random variable. The covariance is $cov(\tilde{x}) = cov(\tau\epsilon + C\Delta\psi) = \tau^2 I + \sigma^2 CC^T$. $\qquad\square$

Proposition A.7 (Fractional Denoising Score Matching)**.** *If $p(\tilde{x}|x_a) \sim \mathcal{N}(x_a, \tau^2 I_{3N})$ and $p(x_a|x_i)$ can be arbitrary distribution, we have*

$$
\begin{aligned}
&E_{p(\tilde{x}|x_a)p(x_a|x_i)p(x_i)}\|GNN_\theta(\tilde{x}) - (\tilde{x} - x_a)\|^2 \\
&\simeq E_{p(\tilde{x})}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x})\|^2 \\
&\simeq E_{p(\tilde{x},x_i)}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}|x_i)\|^2
\end{aligned}
\tag{20}
$$

$\simeq$ *denotes the equivalence as optimization objectives. $\nabla_{\tilde{x}} \log p(\tilde{x}) = -\nabla_{\tilde{x}} E(\tilde{x})$ is the force field of the hybrid noise, because $p(\tilde{x}) = \sum_{i=1}^{n} p(\tilde{x}|x_i)p_0(x_i)$ and $p(\tilde{x}|x_i)$ is given by hybrid noise and with an anisotropic covariance. We add the second equivalence in* (20) *to emphasize the anisotropic force field and that it is not equivalent to directly denoise the hybrid noise $\nabla_{\tilde{x}} \log p(\tilde{x}|x_i) = \Sigma_{\sigma,\tau}^{-1}(x_i - \tilde{x})$, where the coefficient $\Sigma_{\sigma,\tau}^{-1}$ cannot be absorbed in the GNN.*

Proof.

$$
\begin{aligned}
&E_{p(\tilde{x},x_i)}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}|x_i)\|^2 \\
&= E_{p(\tilde{x})}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x})\|^2 + T_1 \\
&= E_{p(\tilde{x},x_a)}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}|x_a)\|^2 + T_1 + T_2 \\
&= E_{p(\tilde{x},x_a,x_i)}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}|x_a)\|^2 + T_1 + T_2 \\
&= E_{p(\tilde{x}|x_a)p(x_a|x_i)p(x_i)}\|GNN_\theta(\tilde{x}) - \frac{x_a - \tilde{x}}{\tau^2})\|^2 + T_1 + T_2.
\end{aligned}
\tag{21}
$$

The first two equations use the result in (Vincent, 2011) (It is proved in Proposition A.8), where T_1, T_2 represents terms do not contain θ. The third equation holds because the part in the expectation does not contain x_i. The fourth equation holds because $-\frac{1}{\tau^2}$ can be absorbed into GNN_θ and $x_i \to x_a \to \tilde{x}$ is a Markov chain, i.e. $p(\tilde{x}, x_a, x_i) = p(\tilde{x}|x_a)p(x_a|x_i)p(x_i)$. $\qquad\square$

Proposition A.8 ((Vincent, 2011) The equivalence between score matching and conditional score matching)**.** *The two minimization objectives below are equivalent, i.e. $J_1(\theta) \simeq J_2(\theta)$.*

$$
J_1(\theta) = E_{p(\tilde{x})}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x})\|^2
\tag{22}
$$

$$
J_2(\theta) = E_{p(\tilde{x}|x)p(x)}\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}|x)\|^2
\tag{23}
$$

Proof. We first expand the square term and observe:

$$
J_1(\theta) = E_{p(\tilde{x})}[\|GNN_\theta(\tilde{x})\|^2] - 2E_{p(\tilde{x})}[\langle GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \log p(\tilde{x})\rangle] + T_3
\tag{24}
$$

$$
J_2(\theta) = E_{p(\tilde{x}|x)p(x)}[\|GNN_\theta(\tilde{x})\|^2] - 2E_{p(\tilde{x}|x)p(x)}[\langle GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \log p(\tilde{x}|x)\rangle] + T_4,
\tag{25}
$$

where T_3, T_4 are constants independent of θ. Therefore, it suffices to show that the middle terms on the right hand side are

16

equal.

$$
\begin{aligned}
&E_{p(\tilde{x})}[\langle GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \log p(\tilde{x}) \rangle] \\
&= \int_{\tilde{x}} p(\tilde{x}) \langle GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \log p(\tilde{x}) \rangle \, \mathrm{d}\tilde{x} \\
&= \int_{\tilde{x}} p(\tilde{x}) \left\langle GNN_\theta(\tilde{x}), \frac{\nabla_{\tilde{x}} p(\tilde{x})}{p(\tilde{x})} \right\rangle \mathrm{d}\tilde{x} \\
&= \int_{\tilde{x}} \langle GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} p(\tilde{x}) \rangle \, \mathrm{d}\tilde{x} \\
&= \int_{\tilde{x}} \left\langle GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \left(\int_x p(\tilde{x}|x) p(x) \mathrm{d}x \right) \right\rangle \mathrm{d}\tilde{x} \\
&= \int_{\tilde{x}} \left\langle GNN_\theta(\tilde{x}), \int_x p(x) \nabla_{\tilde{x}} p(\tilde{x}|x) \mathrm{d}x \right\rangle \mathrm{d}\tilde{x} \\
&= \int_{\tilde{x}} \left\langle GNN_\theta(\tilde{x}), \int_x p(\tilde{x}|x) p(x) \nabla_{\tilde{x}} \log p(\tilde{x}|x) \mathrm{d}x \right\rangle \mathrm{d}\tilde{x} \\
&= \int_{\tilde{x}} \int_x p(\tilde{x}|x) p(x) \langle GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \log p(\tilde{x}|x) \rangle \, \mathrm{d}x \mathrm{d}\tilde{x} \\
&= E_{p(\tilde{x},x)}[\langle GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \log p(\tilde{x}|x) \rangle]
\end{aligned}
\tag{26}
$$

$\square$

B. Experimental Details

B.1. How Does Perturbation Scale Affect the Performance of Two Denoising Methods?

We have discussed in section 3.1.1 that structures sampled by adding coordinate noise can violate the chemical constraints as the noise scale increases, and hybrid noise can alleviate the problem. An empirical verification is provided in Table 6.

Table 6. The average preformance(MAE) of coordinate denoising and Frad with distinct noise scale on seven energy prediction tasks in QM9. We also report the perturbation scale defined by mean absolute coordinate changes caused by applying different noise scales. Frad can achieve better performance than coordinate denoise settings even with large perturbation scale.

	$\tau = 0.004$	$\tau = 0.04$	$\tau = 0.4$	$\sigma = 1, \tau = 0.04$	$\sigma = 2, \tau = 0.04$	$\sigma = 20, \tau = 0.04$
Perturbation Scale	0.00319	0.0319	0.319	0.0635	0.0952	0.6499
Average Performance(meV)	14.06	12.86	15.26	12.06	11.94	12.30

Instead of using the variance of Gaussian, we introduce the perturbation scale defined by mean absolute coordinate changes of all the atoms after applying the noise, denoted by PS. In this way, we can fairly compare the noise scale of distinct noise types. Taking the units and value scale into account, we select seven energy prediction tasks, i.e . $\epsilon_{HOMO}, \epsilon_{LUMO}, \Delta\epsilon, U_0, U, H, G$ in QM9. The results support our motivation as follows.

- Is the challenge of low sampling coverage observed in coordinate denoising? Indeed, in three coordinate denoising settings, $\tau = 0.04$ behaves best. Both larger and smaller noise scale degenerate the performance. A large noise scale leads to more irrational noisy samples while a small noise scale results in trivial denoising tasks. This is also reported by Zaidi et al. (2022), tuning $\tau = 0.04$ as the best hyperparameter.

- Does hybrid noise alleviate the low sampling coverage problem? Yes, the perturbation scale of the hybrid noise can be large without losing the competence. Even with $\sigma = 20$, reflecting notable rotation of the single bonds, it is still superior than all the coordinate denoising settings on average.

- Are better approximations of force field helpful? Yes, it can be inferred from the comparison between ($\tau = 0.04$) and ($\sigma = 2$ or 1, $\tau = 0.04$). Note that the three settings share similar small perturbation scale, suggesting they all possess meaningful samples and non-trivial reconstructing tasks. By contrast, Frad gains further improvements over coordinate denoising, indicating better approximation of force field helps molecular representation learning.

B.2. Does Hybrid Noise Better Approximate Force Field?

In section 3.1.2, we reveal that the distribution of conformation with hybrid noise can capture the anisotropy of molecular force field. In this section, we quantitatively investigate the force field estimation and support our theoretical statements.

Note that both σ and τ are vital for force field estimation. To avoid tuning two parameters at the same time, we evaluate the estimation accuracy by Pearson correlation coefficient ρ between the estimated force field and ground truth, so that the ratio of σ and τ counts rather than the absolute values. In Table 7, we employ the force field predicted by sGDML (Chmiela et al., 2019) as ground truth. For fair comparison, we decouple the sampling and force field calculation. The samples are drawn by perturbing the equilibrium with noise setting ($\tau = 0.04$), ($\sigma = 1$, $\tau = 0.04$), ($\sigma = 20$, $\tau = 0.04$), representing samples from near to far from the equilibrium. As for force field calculation, Table 1 offers the conditional score function of the conformation distribution under distinct noise types. Since we only include one equilibrium $n = 1$, as is discussed in section 2, the probability $p_0 = 1$. Thus the conditional score function equals score function and is exactly the force field under Boltzmann distribution. Note that all the quantities in Table 1 can be specified, except the matrix C in Σ_σ and $\Sigma_{\sigma,\tau}$. Given the equilibrium, C is determined and can be computed by least squares estimation. We utilize the normalized residual sum of squares $C_{error} = \frac{\|\Delta x - C\Delta\psi\|}{\|\Delta x\|}$ to measure the error of the least squares estimation of C. The results are shown in Table 7.

Table 7. The Pearson correlation coefficient between the force field estimation and ground truth.

| Sampling setting | $\tau = 0.04$ | | $\sigma = 1, \tau = 0.04$ | | $\sigma = 20, \tau = 0.04$ | |
| Sample size | 1000 | | 1000 | | 1000 | |
Estimation setting	ρ	C_{error}	ρ	C_{error}	ρ	C_{error}
$\tau = 0.04$	0.5775	-	0.5565	-	0.12047	-
$\sigma = 0.01, \tau = 0.04$	0.57755	0.0003	0.5565	0.00037	0.1205	0.0003
$\sigma = 0.1, \tau = 0.04$	0.578	0.0012	0.558108	0.0012	0.12326	0.00122
$\sigma = 0.5, \tau = 0.04$	0.58459	0.0062	0.5776	0.0062	0.17905	0.00639
$\sigma = 1, \tau = 0.04$	0.5893	0.0126	0.59032	0.0129	0.27628	0.0126
$\sigma = 2, \tau = 0.04$	0.5915	0.02438	0.59544	0.0261	0.37443	0.02514
$\sigma = 20, \tau = 0.04$	0.5921	0.2539	0.5964	0.2506	0.41103	0.2561
$\sigma = 50, \tau = 0.04$	0.5927	0.6052	0.5950	0.6144	0.3964	0.6044
$\sigma = 100, \tau = 0.04$	0.5916	0.9432	0.58728	0.9352	0.2849	0.9298

The findings are threefold. 1. In all sampling settings, hybrid noise achieve better estimation accuracy than coordinate noise. Specifically, σ around 20 and $\tau = 0.04$ best fit the ground truth force field. 2. The accuracy gap between the hybrid noise and the coordinate noise becomes more evident when more samples far from equilibrium are included. These two findings demonstrate the superiority of hybrid noise over coordinate noise. 3. C_{error} remains small in the settings with little angle noise scale, confirming the correctness of Proposition 3.1. When $\sigma \le 20$, C_{error} grows large, indicating an inaccuracy force field calculated by Table 1. As a consequence, we choose $\sigma = 2$, $\tau = 0.04$ as the hyperparameter of Frad.

B.3. Does Learning Force Field Help the Downstream Tasks?

To verify whether learning a force field can improve downstream tasks, we conduct the following experiment: as obtaining the true force field label can be time-consuming, we randomly select 10,000 molecules with fewer than 30 atoms from the PCQM4Mv2 dataset and calculate their precise force field label using DFT. We then pre-train the model by predicting these force field labels obtained from DFT, followed by fine-tuning the model on various sub-tasks from the QM9 and MD17 datasets. Table 8 summarizes the results, which demonstrate that learning the force field improves the performance of downstream tasks compared to training from scratch. These findings suggest that pre-training with learning force field is effective.

B.4. Pseudocode For Pre-training And Fine-tuning Algorithms

In this section, we present pseudocode to elucidate the algorithms described in the paper. Specifically, Algorithm 1 showcases pre-training algorithm of Frad; While the Noisy Nodes, which is the fine-tuning algorithm used for QM9 dataset, is demonstrated in Algorithm 2; Lastly, Algorithm 3 provides the pseudocode for the fine-tuning algorithm used on the MD17 dataset.

Algorithm 1 Applying Frad to pre-training

Require:

    τ: Scale of coordinate noise

    σ: Scale of dihedral angle noise

    GNN_θ: Graph Neural Network with parameter θ

    X: Unlabeled pre-training dataset

    x_i: Input conformation

    T: Training steps

    $\mathcal{N}$: Gaussian distribution

0: **while** $T \neq 0$ **do**

0:    x_i = dataloader(X) {random sample x_i from X}

0:    Find dihedral angles of x_i denoted as $\psi = (\psi_1, ..., \psi_m) \in [0, 2\pi)^m$

0:    Change coordinates of x_i to x_a to satisfy that: $\psi_a = \psi + \Delta\psi$, where ψ_a represents dihedral angles of x_a and $\Delta\psi \sim \mathcal{N}(0, \sigma^2 I_m)$

0:    $\tilde{x} = x_a + \Delta x_i$, where $\Delta x_i \sim \mathcal{N}(0, \tau^2 I_{3N})$, N is atom number of x_i

0:    $\Delta x_i^{pred} = GNN_\theta(\tilde{x})$

0:    Loss = $\|\Delta x_i^{pred} - \Delta x_i\|_2^2$

0:    Optimise(Loss)

0:    $T = T - 1$

0: **end while**=0

Algorithm 2 Noisy Nodes algorithm

Require:

    τ: Scale of coordinate noise

    GNN_θ: Graph Neural Network with parameter θ

    $\text{NoiseHead}_{\theta_n}$: Network module with parameter θ_n for prediction of node-level noise of each atom

    $\text{LabelHead}_{\theta_l}$: Network module with parameter θ_l for prediction of graph-level label of x_i

    X: Training dataset

    x_i: Input conformation

    y_i: Label of x_i

    T: Training steps

    $\mathcal{N}$: Gaussian distribution

    λ_p: Loss weight of property prediction loss

    λ_n: Loss weight of Noisy Nodes loss

0: **while** $T \neq 0$ **do**

0:    x_i, y_i = dataloader(X) {random sample x_i and corresponding label y_i from X}

0:    $\tilde{x} = x_i + \Delta x_i$, where $\Delta x_i \sim \mathcal{N}(0, \tau^2 I_{3N})$, N is atom number of x_i

0:    $y_i^{pred} = \text{LabelHead}_{\theta_l}(GNN_\theta(\tilde{x}))$

0:    $\Delta x_i^{pred} = \text{NoiseHead}_{\theta_n}(GNN_\theta(\tilde{x}))$

0:    Loss = $\lambda_p \text{PropertyPredictionLoss}(y_i^{pred}, y_i) + \lambda_n \|\Delta x_i^{pred} - \Delta x_i\|_2^2$

0:    Optimise(Loss)

0:    $T = T - 1$

0: **end while**=0

Table 8. The performance (MAE) comparison between pre-training with learning force field and training from scratch on 3 sub-tasks from QM9 and MD17 datasets. The top results are in bold.

	QM9:ϵ_{HOMO} (meV)	QM9:ϵ_{LUMO} (meV)	MD17:Aspirin (Force)
training from scratch	19.2	20.9	0.253
pre-training with learning force field	**17.1**	**19.6**	**0.236**

Algorithm 3 Applying Frad to fine-tuning on MD17

Require:
  τ: Scale of coordinate noise
  σ: Scale of dihedral angle noise
  GNN_θ: Graph Neural Network with parameter θ
  $\text{NoiseHead}_{\theta_n}$: Network module with parameter θ_n for prediction of node-level noise of each atom
  $\text{LabelHead}_{\theta_l}$: Network module with parameter θ_l for prediction of graph-level label of x_i
  X: Training dataset
  x_i: Input conformation
  y_i: Label of x_i
  T: Training steps
  $\mathcal{N}$: Gaussian distribution
  λ_p: Loss weight of property prediction loss
  λ_n: Loss weight of Noisy Nodes loss
0: **while** $T \neq 0$ **do**
0:     $x_i, y_i = \text{dataloader}(X)$ {random sample x_i and corresponding label y_i from X}
0:     Find dihedral angles of x_i denoted as $\psi = (\psi_1, ..., \psi_m) \in [0, 2\pi)^m$
0:     Change coordinates of x_i to x_a to satisfy that: $\psi_a = \psi + \Delta\psi$, where ψ_a represents dihedral angles of x_a and $\Delta\psi \sim \mathcal{N}(0, \sigma^2 I_m)$
0:     $\tilde{x} = x_a + \Delta x_i$, where $\Delta x_i \sim \mathcal{N}(0, \tau^2 I_{3N})$, N is atom number of x_i
0:     $y_i^{pred} = \text{LabelHead}_{\theta_l}(GNN_\theta(x_i))$
0:     $\Delta x_i^{pred} = \text{NoiseHead}_{\theta_n} GNN_\theta(\tilde{x})$
0:     $\text{Loss} = \lambda_p \text{PropertyPredictionLoss}(y_i^{pred}, y_i) + \lambda_n \|\Delta x_i^{pred} - \Delta x_i\|_2^2$
0:     Optimise(Loss)
0:     $T = T - 1$
0: **end while**=0

B.5. Hyper-parameter Settings

Hyperparameters for pre-training are listed in Table 9. Details about Learning rate decay policy can be refered in https://hasty.ai/docs/mp-wiki/scheduler/reducelronplateaustrong-reducelronplateau-explained-strong. When searching rotatable bonds, the hydrogen atoms are taken into account, which can further improve the performance.

Hyperparameters for fine-tuning on MD17 are listed in Table 10. We test our model in two ways of data splitting. Correspondingly, there are two batch sizes proportional to the training data size.

Hyperparameters for fine-tuning on QM9 are listed in Table 11. The cosine cycle length is set to be 500000 for α, $ZPVE$, U_0, U, H, G and 300000 for other tasks for fully converge. Notice that because the performance of QM9 and MD17 is quite stable for random seed, we will not run cross-validation. This also follows the main literature (Schütt et al., 2018; 2021; Liu et al., 2022b;a).

Table 9. Hyperparameters for pre-training.

Parameter	Value or description
Train Dataset	PCQM4MV2
Batch size	70
Optimizer	AdamW
Warm up steps	10000
Max Learning rate	0.0004
Learning rate decay policy	Cosine
Learning rate factor	0.8
Cosine cycle length	400000
Network structure	Keep aligned with downstream settings respectively on QM9 and MD17
Dihedral angle noise scale(type: Gaussian)	2
Coordinate noise scale(type: Gaussian)	0.04

Table 10. Hyperparameters for fine-tuning on MD17.

Parameter	Value or description
Train/Val/Test Splitting*	9500/500/remaining data (950/50/remaining data)
Batch size*	80 (8)
Optimizer	AdamW
Warm up steps	1000
Max Learning rate	0.001
Learning rate decay policy	ReduceLROnPlateau (Reduce Learning Rate on Plateau) scheduler
Learning rate factor	0.8
Patience	30
Min learning rate	1.00E-07
Network structure	TorchMD-NET
Head number	8
Layer number	6
RBF number	32
Activation function	SiLU
Embedding dimension	128
Force weight	0.8
Energy weight	0.2
Noisy Nodes denoise weight	0.1
Dihedral angle noise scale(type: Gaussian)	20
Coordinate noise scale(type: Gaussian)	0.005

C. More Preliminaries

C.1. Boltzmann Distribution

From the prior knowledge in statistical physics, the conformations of a molecule can be viewed as in Boltzmann distribution $p_{physical}(\tilde{x}) = \frac{1}{Z} e^{-\frac{E_{physical}(\tilde{x})}{k_B T}}$ (Boltzmann, 1868), where $E(\tilde{x})$ is the (potential) energy function, $\tilde{x} \in \mathbb{R}^{3N}$ is the position of the atoms, i.e. conformation, N is the number of atoms in the molecule, T is the temperature, k_B is the Boltzmann constant and Z is the normalization factor. When employing neural networks to fit the energy function or its gradient, the constant $k_B T$ can be absorbed in the energy function, i.e. $p_{physical}(\tilde{x}) \propto exp(-E_{physical}(\tilde{x}))$.

Table 11. Hyperparameters for fine-tuning on QM9.

Parameter	Value or description
Train/Val/Test Splitting	110000/10000/remaining data
Batch size	128
Optimizer	AdamW
Warm up steps	10000
Max Learning rate	0.0004
Learning rate decay policy	Cosine
Learning rate factor	0.8
Cosine cycle length*	300000 (500000)
Network structure	TorchMD-NET
Head number	8
Layer number	8
RBF number	64
Activation function	SiLU
Embedding dimension	256
Head Standardize AtomRef	Applied according to https://github.com/torchmd/torchmd-net/issues/64
Label weight	1
Noisy Nodes denoise weight	0.1(0.2)
Coordinate noise scale(type: Gaussian)	0.005

C.2. Molecular Force Field Learning

Here we reveal the connection between denoising and force field learning. Under the Boltzmann distribution assumption, the score function of the conformation distribution is the molecular force field.

$$\nabla_{\tilde{x}} \log p(\tilde{x}) = -\nabla_{\tilde{x}} E(\tilde{x}), \tag{27}$$

where $-\nabla_x E(\tilde{x})$ is referred to as the molecular force field, indicating the force on each atom. When the variance of the Gaussian distribution is isotropic diagonal $\Sigma = \tau^2 I_{3N}$, the conditional score function is given by

$$\nabla_{\tilde{x}} \log p(\tilde{x}|x_i) = -\frac{\tilde{x} - x_i}{\tau^2}. \tag{28}$$

Then, by the fact that score matching is equivalent to conditional score matching (Vincent, 2011) (It is proved in Proposition A.8.), we establish the equivalence between denoising and force field learning.

$$
\begin{aligned}
& E_{p(\tilde{x})} \|GNN_\theta(\tilde{x}) - (-\nabla_{\tilde{x}} E(\tilde{x}))\|^2 \\
= & E_{p(\tilde{x})} \|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x})\|^2 \\
= & E_{p(\tilde{x},x_i)} \|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}|x_i)\|^2 + T \\
= & E_{p(\tilde{x},x_i)} \|GNN_\theta(\tilde{x}) - \frac{x_i - \tilde{x}}{\tau^2})\|^2 + T,
\end{aligned}
\tag{29}
$$

where $GNN_\theta(\tilde{x})$ denotes a graph neural network with parameters θ which takes conformation $\tilde{x}$ as an input and returns node-level predictions, T represents terms independent of θ. Therefore, $\arg\min_\theta E_{p(\tilde{x},x_i)} \|GNN_\theta(\tilde{x}) - \frac{x_i - \tilde{x}}{\tau^2})\|^2 = \arg\min_\theta E_{p(\tilde{x})} \|GNN_\theta(\tilde{x}) - (-\nabla_{\tilde{x}} E(\tilde{x}))\|^2$.

The coefficient $-\frac{1}{\tau^2}$ are constant that do not rely on the input $\tilde{x}$, so it can be "absorbed" into GNN_θ (Zaidi et al., 2022). This is because we can define $GNN_{\theta'} = -\tau^2 GNN_\theta$, where the parameters in the last layer (linear transformation layer) of $GNN_{\theta'}$ is $-\tau^2$ times that of GNN_θ and other parameters remain the same. Once one of the GNN is optimized, the optimal parameters of the other can also be determined, so these two optimization goals can be regarded as equivalent. Another understanding is that the two GNN learn the same force field label but up to a unit conversion. In conclusion, typical

denoising loss and force field fitting loss are equivalent optimization objectives.

$$\min_{\theta} E_{p(\tilde{x},x_i)}\|GNN_{\theta}(\tilde{x}) - (\tilde{x} - x_i)\|^2$$
$$\simeq \min_{\theta} E_{p(\tilde{x})}\|GNN_{\theta}(\tilde{x}) - (-\nabla_{\tilde{x}}E(\tilde{x}))\|^2. \tag{30}$$

It is assumed in literature (Zaidi et al., 2022; Jiao et al., 2022) that learning the molecular force field gives rise to useful representations for downstream tasks. This is because in computational chemistry, force field and potential energy are fundamental physical quantities that depend on molecular conformation, capturing the interactions between atoms in 3D space. In addition, labels for many downstream tasks can be calculated by energy or force field based method such as density functional theory (DFT) and molecular dynamics, showing the close relationship between downstream tasks and force field (Chmiela et al., 2017). Empirically, Liu et al. (2022a); Luo et al. (2022) demonstrate that learning the energy of the input conformation helps molecular representation learning. Also, learning a molecular force field is already revealed to be beneficial for molecular property prediction (Zaidi et al., 2022; Jiao et al., 2022; Liu et al., 2022a). As a consequence, we take force field as a reasonable pre-training objective and try to encode the information of energy landscape into our model. Since for most downstream tasks, related structures have relatively low energy, we distill the aim of denoising into learning a force field as accurately as possible for common molecules.

D. More Related Work

D.1. A Brief Overview of Molecular Pre-training

Pre-training is an important approach for molecular representation learning. Traditionally, molecular pre-training concentrates on 1D SMILES strings (Wang et al., 2019; Honda et al., 2019; Chithrananda et al., 2020; Zhang et al., 2021a; Xue et al., 2021; Guo et al., 2022) and 2D molecular graphs (Rong et al., 2020; Li et al., 2020; Zhang et al., 2021b; Li et al., 2021; Zhu et al., 2021; Wang et al., 2022b;c; Fang et al., 2022b; Lin et al., 2022). Inspired by the pre-training methods in CV and NLP, they implement masking and contrastive self-supervised learning tasks to improve molecular representations for 1D and 2D inputs.

Recently, more methods try to incorporate 3D atomic coordinate position as inputs, which is a more informative and physically intrinsic representation for molecules. Earlier methods utilize 3D information as a supplement of 2D input and learn the representation on 2D graphs in a contrastive or generative way (Liu et al., 2021; Li et al., 2022; Zhu et al., 2022; Stärk et al., 2022a). Afterwards, recent methods develop self-supervised learning tasks specifically for 3D geometry data and learn the representation directly on 3D inputs (Fang et al., 2022a; Zhou et al., 2022; Luo et al., 2022; Zaidi et al., 2022; Liu et al., 2022a; Jiao et al., 2022).

So far, three kinds of pre-training tasks tailored for 3D structures have been designed, including denoising, geometry masking and geometry prediction. As for masking, Fang et al. (2022a) masks and predicts the bond lengths and bond angles in molecular structures, whereas Zhou et al. (2022) mask and predict atom types based on noisy geometry. With respect to predictive methods, Fang et al. (2022a) propose an atomic distance prediction task with bond lengths and bond angles as inputs. Though it aims to capture global spatial structures, the distance prediction task is partly trivial because some atomic distances can be easily calculated by bond lengths and angles. Moreover, coordinate denoising enjoys a force field interpretation and thus helps downstream tasks. Therefore, we focus on coordinate denoising method in our work.

Typical denoising task is corrupting and reconstructing 3D coordinates of the molecule (Zaidi et al., 2022; Luo et al., 2022; Zhou et al., 2022). To be specific, firstly the input molecular structure (usually equilibrium) is perturbed by adding i.i.d. Gaussian noise to its atomic coordinates and then the model is trained to predict the noise from the corrupted structure. Based on coordinate denoising and its molecular force field interpretation, some works introduce equivariance to the molecular energy function. Jiao et al. (2022) design a Riemann-Gaussian noise so that the energy function is E(3)-invariant. Likewise, in order to maintain SE(3)-invariance during coordinate denoising, Liu et al. (2022a) propose denoise on pairwise atomic distance. Note that our work designs a novel approach that captures the anisotropic nature of molecular energy function, which is neglected by the existing denoising methods.

D.2. Denoising and Score Matching

Using noise to improve the generalization ability of neural networks has a long history(Sietsma & Dow, 1991; Bishop, 1995). Denoising autoencoders(Vincent et al., 2008; 2010) propose a denoising strategy to learn robust and effective representations.

They interpret denoising as a way to define and learn the data manifold. As for GNNs, Hu et al. (2019); Kong et al. (2020); Sato et al. (2021); Godwin et al. (2021) demonstrate that training with noisy inputs can improve performance. Specifically, Noisy Nodes (Godwin et al., 2021) implement denoising as an auxiliary loss to relieve over-smoothing and help molecular property prediction.

Score matching is an energy-based generative model to maximum likelihood for unnormalized probability density models whose partition function is intractable. Denoising is proved to be closely related to score-matching when the noise is standard gaussian (Vincent, 2011). This is successfully applied in generative modelling (Song & Ermon, 2019; 2020; Hu et al., 2019; Hoogeboom et al., 2023) and energy-based molecule modelling to learn a force field (Zaidi et al., 2022; Jiao et al., 2022; Liu et al., 2022a). Though generative models and force field learning both rely on the result of (Vincent, 2011), they are different in assumptions and aims in practice.