Coordinate Denoising for Non-Equilibrium Molecular Representation Learning

Qianwei Tang[1,2] Baile Xu[1,2†] Jian Zhao[1,3] Furao Shen[1,2]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
[2]School of Artificial Intelligence, Nanjing University, China
[3]School of Electronic Science and Engineering, Nanjing University, China
qweit@smail.nju.edu.cn {xubaile, jianzhao, frshen}@nju.edu.cn

Abstract

Three-dimensional molecular representation learning has shown great promise in modeling chemical structures and their properties. However, most existing approaches implicitly assume molecules are at or near equilibrium states. This assumption breaks down for non-equilibrium structures—ubiquitous in molecular dynamics (MD) trajectories—where standard coordinate denoising techniques fail because the direct equivalence between denoising scores and atomic forces no longer holds. To bridge this gap, we propose Node Denoising on non-Equilibrium Molecules (NDeM), a novel auxiliary task grounded in a second-order finite difference approximation of the potential energy surface. By explicitly accounting for the non-zero inherent forces in non-equilibrium states, NDeM provides a theoretically sound denoising objective applicable to arbitrary molecular conformations. Crucially, our method is designed as a lightweight, architecture-agnostic plugin that requires no pre-training and can be seamlessly integrated into various supervised learning pipelines. Extensive experiments across diverse benchmarks, including MD17, QM9, and the large-scale OC20 dataset, demonstrate that NDeM consistently improves baseline models, yielding highly competitive performance and validating its robustness across both equilibrium and non-equilibrium regimes.

1. Introduction

Molecular representation learning is a cornerstone of modern computational chemistry, driving progress in tasks such as de novo molecular generation [2, 20], drug-drug interaction prediction [1], and protein-ligand binding affinity estimation [13, 46]. Early methods, inspired by NLP and computer vision [6, 9, 37], often represented molecules as 1D SMILES strings [43, 45] or 2D graphs [25, 42]. However, these simplified forms overlook critical 3D geometric information—bond lengths, angles, and torsional conformations—that dictates a molecule's physicochemical state and energy profile. Consequently, research has shifted toward 3D molecular representation learning [38] to explicitly encode fine-grained stereostructural features.

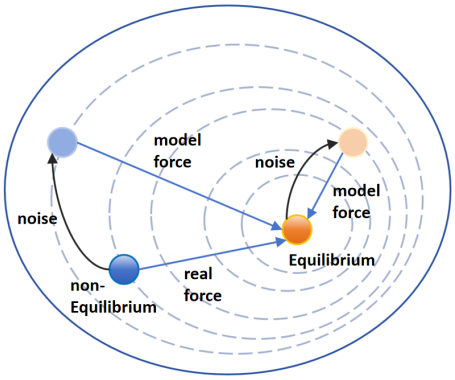

Figure 1. Denoising Equilibrium vs. Non-Equilibrium Molecules. For equilibrium molecules (right), predicting noise is equivalent to learning molecular forces. For non-equilibrium molecules (left), inherent non-zero forces confound this relationship. NDeM addresses this by explicitly accounting for these real forces in the denoising objective.

To respect physical laws like roto-translational invariance, physically-aligned paradigms such as coordinate denoising [10, 44, 47] have emerged as powerful self-supervised approaches. By training networks to recover original structures from Gaussian-perturbed coordinates, this process becomes theoretically equivalent to Score Matching—effectively learning the potential energy surface (PES) gradient, which corresponds to the force field at equilibrium [44]. Mastering these fundamental forces and the PES [8, 19] significantly boosts performance across diverse downstream tasks.

Despite its immense success, this standard denoising paradigm is implicitly predicated on equilibrium-state theory. It tacitly assumes that the unperturbed "clean"

†Corresponding author.

molecule lies at a local energy minimum, where the inherent net atomic forces are effectively zero. This assumption, however, fundamentally breaks down for non-equilibrium structures—which are ubiquitous in realistic scenarios such as Molecular Dynamics (MD) simulations of chemical reactions, catalysis, or protein folding. In these transient states, atoms experience significant inherent forces driving them toward lower energy configurations [24]. As visually illustrated in Figure 1, simply predicting the artificially added noise in these states fails to accurately reflect the true underlying force field, because the training objective becomes confounded by the molecule's inherent, non-zero dynamics. Consequently, applying conventional denoising methods to general, non-equilibrium molecular data often yields suboptimal representations that lack robustness and fail to capture the true complexity of the PES.

To bridge this critical gap, we propose a novel, generalized training framework grounded in a rigorous finite-difference approximation of the potential energy surface. By augmenting classical supervised force field learning with a physically-motivated auxiliary task based on second-order finite differences, we effectively decouple the inherent molecular forces from the artificially injected noise. This novel formulation provides a theoretically sound denoising objective that is strictly valid for both equilibrium states (as a special case where inherent forces are zero) and general non-equilibrium states, enabling more precise and uniform learning of the global PES. Importantly, unlike many prior methods that require complex pre-training stages, our method is designed as a lightweight, architecture-agnostic auxiliary task that can be seamlessly integrated into virtually any existing supervised training pipeline for energy and force prediction.

In summary, we introduce Node Denoising on non-Equilibrium Molecules (NDeM), an efficient, generalizable, and theoretically-grounded training method. Our primary contributions are threefold:

- We derive a rigorously physically-motivated denoising objective specifically tailored for general non-equilibrium molecules, based on a second-order Taylor expansion of the potential energy, providing a clear theoretical foundation for generalized denoising beyond equilibrium states.
- We formulate NDeM as a simple, flexible, and architecture-agnostic auxiliary task that can be readily integrated into standard force and energy prediction models without requiring separate pre-training or complex architectural modifications.
- We conduct extensive experiments on standard benchmarks, including small molecules (MD17), equilibrium properties (QM9), and large-scale catalytic systems (OC20), demonstrating that NDeM consistently enhances the performance and robustness of baseline models across diverse molecular states and tasks.

2. Related Work

2.1. Denoising and Score Matching in Molecules

While noise injection is a standard regularization technique [3, 36], molecular denoising uniquely connects to physical laws. Noisy Nodes [17] first used it as an auxiliary task for property prediction. Later breakthroughs established its equivalence to score matching [18, 41], enabling force field learning via denoising [10, 44]. However, these methods assume zero inherent forces, failing for non-equilibrium MD trajectories. Despite recent interest in broader generalization [23, 24], rigorous non-equilibrium formulations remain scarce. Our work directly addresses this gap by extending the denoising paradigm to non-equilibrium molecules via a physically-grounded finite-difference approach.

2.2. Equivariant Molecular Architectures

Effective 3D molecular modeling necessitates roto-translational inductive biases, typically implemented via equivariant neural networks (ENNs). Pioneering architectures like TFN [40] and SE(3)-Transformer [12] leveraged irreducible representations (irreps) and tensor products to achieve SE(3)-equivariance. Subsequent research has focused on improving expressivity through non-linear message passing (e.g., SEGNN [5], Equiformer [22]) or enhancing efficiency by simplifying SO(3) convolutions (e.g., eSCN [30], EquiformerV2 [23]). Despite these architectural strides, most methods still rely on standard energy/force regression as the primary training objective. Our work provides an orthogonal contribution: we propose a generalizable auxiliary learning objective that can be seamlessly integrated with any such equivariant backbone to further boost performance, particularly on complex non-equilibrium data.

3. Preliminary

This section reviews the theoretical foundations linking molecular coordinate denoising to physical force field learning. Starting from statistical physics principles, we derive how denoising objectives serve as proxies for interatomic forces and highlight the critical assumptions underpinning this equivalence.

3.1. Boltzmann Distribution and Molecular Forces

From the perspective of statistical mechanics, a molecular system in thermodynamic equilibrium at temperature T follows the canonical ensemble. The probability density of observing a specific molecular conformation $\mathbf{x} \in \mathbb{R}^{3N}$ (where N is the number of atoms and $3N$ represents their Cartesian coordinates) is governed by the Boltzmann distribution [4]:

$$p_{physical}(\mathbf{x}) = \frac{1}{Z} e^{-\frac{E(\mathbf{x})}{k_B T}}, \tag{1}$$

where $E(\mathbf{x})$ is the potential energy surface (PES) of the molecule, k_B is the Boltzmann constant, and $Z = \int e^{-\frac{E(\mathbf{x})}{k_B T}} d\mathbf{x}$ is the partition function, which acts as a normalization constant. A central quantity of interest in molecular dynamics is the atomic force field $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^{3N}$, defined as the negative gradient of the potential energy: $\mathbf{F}(\mathbf{x}) = -\nabla_{\mathbf{x}} E(\mathbf{x})$.

By taking the gradient of the log-likelihood of the Boltzmann distribution, we establish a direct link between the statistical score function $\nabla_{\mathbf{x}} \log p_{physical}(\mathbf{x})$ and the physical force field:

$$\nabla_{\mathbf{x}} \log p_{physical}(\mathbf{x}) = -\frac{1}{k_B T} \nabla_{\mathbf{x}} E(\mathbf{x}) = \frac{1}{k_B T} \mathbf{F}(\mathbf{x}). \tag{2}$$

This fundamental relation implies that learning the score function of the molecular data distribution is theoretically equivalent to learning the molecular force field (up to a constant thermal factor $1/k_B T$, which is often absorbed into the model or set to 1 in neural implementations).

3.2. Gaussian Mixture Approximation

While theoretically elegant, directly learning the score function from data is challenging because the true underlying distribution $p_{physical}(\mathbf{x})$ is unknown and complex. To make progress, we must adopt tractable approximations. For molecules at or near stable equilibrium states, their conformational distribution arises primarily from thermal fluctuations around localized energy minima. Thus, it is standard practice to approximate this distribution using a mixture of Gaussians centered at these discrete equilibrium conformations $\{\mathbf{x}_i\}_{i=1}^n$ [44]:

$$p_{physical}(\mathbf{x}) \approx p(\mathbf{x}) = \sum_{i=1}^{n} p(\mathbf{x}|\mathbf{x}_i) p_0(\mathbf{x}_i), \tag{3}$$

where $p_0(\mathbf{x}_i)$ is the prior probability of equilibrium state i, and the conditional term $p(\mathbf{x}|\mathbf{x}_i) = \mathcal{N}(\mathbf{x}; \mathbf{x}_i, \Sigma)$ models thermal fluctuations as Gaussian noise with covariance Σ. This Gaussian assumption provides an analytically tractable bridge between data samples and the underlying energy landscape.

3.3. Equivalence Between Denoising and Force Learning

We now formally demonstrate why, under this Gaussian assumption, the pervasive coordinate denoising task is equivalent to learning the molecular force field. Let us assume isotropic thermal noise, $\Sigma = \sigma^2 \mathbf{I}$. For a perturbed conformation $\tilde{\mathbf{x}}$ generated from an equilibrium center $\mathbf{x}_i$ (i.e., $\tilde{\mathbf{x}} = \mathbf{x}_i + \epsilon$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$), the conditional score function is analytically simple:

$$\nabla_{\tilde{\mathbf{x}}} \log p(\tilde{\mathbf{x}}|\mathbf{x}_i) = \nabla_{\tilde{\mathbf{x}}} \left(-\frac{\|\tilde{\mathbf{x}} - \mathbf{x}_i\|^2}{2\sigma^2} \right) = -\frac{\tilde{\mathbf{x}} - \mathbf{x}_i}{\sigma^2} = -\frac{\epsilon}{\sigma^2}. \tag{4}$$

To learn the force field without explicit force labels, we can employ Score Matching (SM) to train a neural network $\mathbf{s}_\theta(\tilde{\mathbf{x}})$ (e.g., an equivariant GNN) to approximate the true data score. The explicit SM objective is generally intractable as it requires the unknown $\nabla_{\tilde{\mathbf{x}}} \log p(\tilde{\mathbf{x}})$. Fortunately, Vincent [41] proved that this can be equivalently solved via Denoising Score Matching (DSM) by conditioning on the clean data $\mathbf{x}_i$:

$$\begin{aligned} \mathcal{L}_{DSM}(\theta) &= \mathbb{E}_{p(\mathbf{x}_i) p(\tilde{\mathbf{x}}|\mathbf{x}_i)} \left[\left\| \mathbf{s}_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log p(\tilde{\mathbf{x}}|\mathbf{x}_i) \right\|^2 \right] \\ &= \mathbb{E}_{p(\mathbf{x}_i) p(\epsilon)} \left[\left\| \mathbf{s}_\theta(\mathbf{x}_i + \epsilon) - \left(-\frac{\epsilon}{\sigma^2} \right) \right\|^2 \right]. \end{aligned} \tag{5}$$

By defining our network output as $\text{GNN}_\theta(\tilde{\mathbf{x}}) = -\sigma^2 \mathbf{s}_\theta(\tilde{\mathbf{x}})$, this objective simplifies exactly to the standard coordinate denoising loss (reconstruction of the noise vector):

$$\mathcal{L}_{\text{denoise}}(\theta) \propto \mathbb{E}_{\mathbf{x}_i, \epsilon} \left[\left\| \text{GNN}_\theta(\tilde{\mathbf{x}}) - \epsilon \right\|^2 \right]. \tag{6}$$

Combining this result with Eq. 2 and Eq. 4 reveals the fundamental equivalence at the core of equilibrium-based pre-training:

$$\underbrace{\text{GNN}_\theta(\tilde{\mathbf{x}}) \approx \epsilon = \mathbf{x}_i - \tilde{\mathbf{x}}}_{\text{Denoising Objective}} \iff \underbrace{\text{GNN}_\theta(\tilde{\mathbf{x}}) \propto -\mathbf{F}(\tilde{\mathbf{x}})}_{\text{Force Learning Objective}}. \tag{7}$$

Eq. 7 states that by learning to denoise (predicting $\mathbf{x}_i - \tilde{\mathbf{x}}$), the network inherently learns the restoring force field $-\mathbf{F}(\tilde{\mathbf{x}})$ that pulls the molecule back to its equilibrium state $\mathbf{x}_i$. Crucially, this powerful equivalence holds only when $\mathbf{x}_i$ is a true equilibrium state (energy minimum). As we will show in Section 4, for non-equilibrium MD snapshots where inherent forces are non-zero, this simple relation breaks down, necessitating our proposed generalized approach.

4. Method

In this section, we first formalize the standard problem setup for molecular energy and force prediction, providing explicit definitions and baseline training objectives. We then analyze the theoretical limitations of classical coordinate denoising when applied to non-equilibrium structures. To fundamentally address these limitations, we derive Node Denoising on non-Equilibrium Molecules (NDeM), grounded in a second-order finite-difference approximation of the potential energy surface. Finally, we detail the practical implementation of NDeM as a seamless auxiliary task within standard supervised training pipelines.

4.1. Problem Setup

The precise computation of quantum mechanical properties, particularly potential energy and atomic forces, is fundamental to understanding molecular systems. We consider a dataset of 3D molecules, where each molecule M consisting of N atoms is represented as an unordered set:

$$M = \{(z_i, \mathbf{x}_i) \mid i \in \{1, \dots, N\}\}, \tag{8}$$

where $z_i \in \mathbb{N}$ denotes the atomic number, and $\mathbf{x}_i \in \mathbb{R}^3$ specifies the 3D spatial coordinate of the i-th atom. The ground-truth quantum chemical state is characterized by a scalar potential energy $E_{gt}(M) \in \mathbb{R}$, and an atomic force vector field $F_{gt}(M) = \{\mathbf{f}_i \in \mathbb{R}^3 \mid i \in \{1, \dots, N\}\}$, where $\mathbf{f}_i = -\nabla_{\mathbf{x}_i} E_{gt}(M)$ is the force acting on the i-th atom.

A molecule is at an *equilibrium* state if it resides at a local potential energy minimum, meaning inherent atomic forces are negligible ($\|F_{gt}(M)\| \approx 0$). In contrast, *non-equilibrium* structures—which dominate molecular dynamics (MD) trajectories—exhibit significant non-zero forces, driving the system towards lower energy states.

We employ an equivariant Graph Neural Network (GNN) to approximate these quantities: $\hat{E} = \text{GNN}_E(M)$ and $\hat{F} = \text{GNN}_F(M) = \{\hat{\mathbf{f}}_i\}_{i=1}^N$. The model is optimized by minimizing a weighted multi-task loss function over the dataset:

$$\begin{aligned}\mathcal{L}_{\text{orig}} &= \lambda_E \mathcal{L}_E + \lambda_F \mathcal{L}_F \\ &= \lambda_E \left| E^* - \hat{E} \right| + \lambda_F \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{f}_i^* - \hat{\mathbf{f}}_i \right\|^2, \end{aligned} \tag{9}$$

where λ_E, λ_F are balancing coefficients. To ensure numerical stability, we regress on normalized targets:

$$E^* = \frac{E_{gt} - \mu_E}{\sigma_E}, \quad \mathbf{f}_i^* = \frac{\mathbf{f}_{gt,i}}{\sigma_F}, \tag{10}$$

where μ_E, σ_E are the mean and standard deviation of energies, and σ_F is the root mean square of force components across the training set.

4.2. Node Denoising on Non-Equilibrium Molecules

4.2.1. Classical Node Denoising and its Limitations

Classical coordinate denoising is often employed as a self-supervised task. Given a molecular conformation $X = \{\mathbf{x}_i\}_{i=1}^N$, a corrupted version $\tilde{X} = \{\tilde{\mathbf{x}}_i\}_{i=1}^N$ is generated by injecting independent Gaussian noise ϵ_i into each atom:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + \epsilon_i, \quad \text{where } \epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{11}$$

A denoising model is trained to reconstruct the noise vector $\epsilon = \{\epsilon_i\}_{i=1}^N$ from $\tilde{X}$ by minimizing:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{p(\tilde{X}|X)} \left[\frac{1}{N} \sum_{i=1}^{N} \left\| \text{GNN}_{\text{denoise}}(\tilde{X})_i - \epsilon_i \right\|^2 \right]. \tag{12}$$

As discussed in Section 3, for equilibrium structures where inherent forces are zero, this objective is equivalent to learning the force field, i.e., $\text{GNN}_{\text{denoise}}(\tilde{X}) \propto -F(\tilde{X})$. However, this equivalence breaks down for *non-equilibrium* structures. When the original structure X has inherent forces $F(X) \neq \mathbf{0}$, simply predicting the added noise ϵ ignores these dynamics. The denoising task becomes an ill-posed proxy for force learning, as it aims to "restore" a transient state X that is not at an energy minimum (see Figure 1).

4.2.2. Finite Difference Formulation

To derive a valid denoising objective for general non-equilibrium states, we examine the local geometry of the Potential Energy Surface (PES) via a second-order Taylor expansion around an arbitrary conformation X:

$$E(\tilde{X}) \approx E(X) + \nabla E(X)^\top \epsilon + \frac{1}{2} \epsilon^\top \nabla^2 E(X) \epsilon, \tag{13}$$

where $\epsilon = \tilde{X} - X$ is the perturbation, $\nabla E(X) = -F(X)$ is the negative force field, and $\nabla^2 E(X) = \mathbf{H}(X)$ is the Hessian matrix capturing PES curvature. Taking the gradient of Eq. 13 with respect to $\tilde{X}$ relates the forces at the two states:

$$\nabla_{\tilde{X}} E(\tilde{X}) \approx \nabla_X E(X) + \mathbf{H}(X)(\tilde{X} - X). \tag{14}$$

Substituting $F = -\nabla E$, we obtain the fundamental relation:

$$-F(\tilde{X}) \approx -F(X) + \mathbf{H}(X)\epsilon. \tag{15}$$

While physically precise, explicitly computing the Hessian $\mathbf{H}(X)$ during training is computationally prohibitive.

Simplification Assumption. We assume that for small local perturbations ϵ, the local PES curvature is approximately isotropic, allowing us to approximate the Hessian with a scalar scaled identity matrix, $\mathbf{H}(X) \approx c\mathbf{I}$, where c is a learnable scalar representing the local curvature of the PES. Under this assumption, Eq. 15 simplifies to a tractable *finite-difference* relation:

$$F(\tilde{X}) - F(X) \approx -c\epsilon. \tag{16}$$

This relation is the core of NDeM. Unlike classical denoising, it explicitly decouples the inherent force $F(X)$ from the perturbation-induced force change $-c\epsilon$. In our implementation, instead of treating c as a fixed hyperparameter, we formulate it as an implicit learnable parameter within the auxiliary force head $GNN_{F'}$. This allows the model to adaptively capture the potential energy surface's curvature across diverse molecular conformations, providing a more flexible and theoretically aligned training objective. For equilibrium states ($F(X) \approx \mathbf{0}$), this recovers classical denoising ($F(\tilde{X}) \approx -c\epsilon$). Crucially, it remains valid for

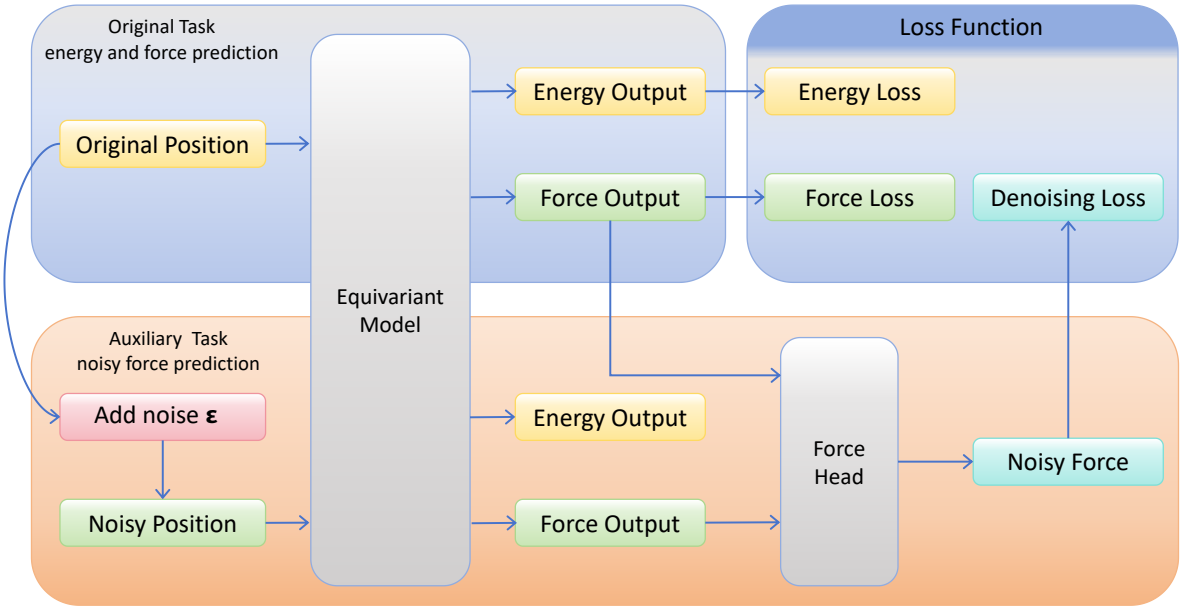

Figure 2. Schematic illustration of Node Denoising as an auxiliary task. The auxiliary task takes the noise-perturbed molecule as input, processes it through the same model, and passes the resulting latent representation through an additional equivariant output head—matching the shape of the molecular force—to obtain the Noisy Force. This is used to compute the denoising loss. Jointly performing both tasks enhances the model's generalization capability for molecular force learning.

non-equilibrium states ($F(X) \neq \mathbf{0}$), providing a generalized training objective that is robust to the inherent dynamics of the molecule.

4.3. Auxiliary Task Implementation

We implement NDeM as an integrated auxiliary task augmenting standard supervised learning, as shown in Figure 2.

4.3.1. NDeM Loss and Total Objective

Based on Eq. 16, we define the NDeM auxiliary loss. To avoid interference between the main task (predicting realistic forces) and the auxiliary task (predicting perturbed forces), we employ a separate auxiliary output head, $\text{GNN}_{F'}$, which shares the equivariant backbone with the main head GNN_F. The NDeM loss enforces finite-difference consistency:

$$\mathcal{L}_{\text{NDeM}} = \mathbb{E}_{X,\epsilon}\left[\frac{1}{N}\sum_{i=1}^{N}\left\|(\hat{\mathbf{f}}_i(X) - \hat{\mathbf{f}}'_i(\tilde{X})) - \epsilon_i^*\right\|^2\right], \tag{17}$$

where $\hat{\mathbf{f}}_i(X)$ is the main head prediction, $\hat{\mathbf{f}}'_i(\tilde{X})$ is the auxiliary head prediction, and ϵ_i^* is the normalized ground-truth noise. The total objective combines both losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{orig}} + \lambda_{\text{NDeM}}\mathcal{L}_{\text{NDeM}}, \tag{18}$$

where λ_{NDeM} weights the auxiliary task. This joint optimization ensures the model learns both absolute force values (via $\mathcal{L}_{\text{orig}}$) and local force gradients (via $\mathcal{L}_{\text{NDeM}}$).

4.3.2. Multi-scale Noise Injection

To balance the validity of the Taylor expansion with exploration, we use a *multi-scale masked noise* strategy [24]. In each iteration, we randomly mask a subset of atoms and only apply noise to them. The noise standard deviation σ is sampled from a range $[\sigma_{\min}, \sigma_{\max}]$ for each batch, encouraging robust representations across different perturbation scales.

4.3.3. Discussion

NDeM offers key advantages: (1) It is strictly grounded in quantum mechanical principles, acting as a physically-informed regularizer. (2) Unlike self-supervised pre-training, it is an efficient auxiliary loss optimized concurrently with the main task. (3) Most importantly, by explicitly accounting for non-equilibrium background forces, it provides a theoretically sound generalization applicable to any molecular dataset, from simple equilibrium geometries to complex reaction trajectories.

Table 1. Comparison of force prediction performance (MAE, lower is better) on the MD17 dataset. Models marked with an asterisk (*) indicate that their results for Benzene are unavailable, and SE(3)-DDM uses a different Benzene dataset version.

	Aspirin	Benzene	Ethanol	Malonaldehyde	Naphthalene	Salicylic Acid	Toluene	Uracil
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
PaiNN*	0.338	–	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
SE(3)-DDM*	0.453	–	0.166	0.288	0.129	0.226	0.122	0.183
Frad	0.2087	0.1994	0.0910	**0.1415**	0.0530	0.1081	0.0540	**0.0760**
SliDe	0.1740	**0.1691**	0.0882	0.1538	0.0483	0.1006	0.0540	0.0825
NDeM	**0.1654**	0.1802	**0.0868**	0.1439	**0.0480**	**0.1004**	**0.0536**	0.0806

Table 2. Comparison of performance (MAE, lower is better) on the QM9 dataset.

	μ (D)	α (a_0^3)	homo (meV)	lumo (meV)	gap (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
DimeNet++	0.030	0.044	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	0.025	0.045	22.8	18.9	31.1	0.27	**1.120**	6.26	6.36	6.33	7.78	0.022
TorchMD-NET	0.011	0.059	20.3	17.5	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	1.22	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
Frad	0.010	0.0374	15.3	13.7	27.8	0.3419	1.418	5.33	5.62	5.55	6.19	0.020
SliDe	0.0087	0.0366	**13.6**	**12.3**	26.2	0.3405	1.521	**4.28**	**4.29**	**4.26**	**5.37**	**0.019**
NDeM	**0.0082**	**0.0356**	**13.6**	**12.3**	**26.0**	0.3127	1.210	5.31	5.34	4.89	5.51	**0.019**

5. Experiment

In this section, we conduct experiments to evaluate our training methodology. We assess our approach on widely-used molecular datasets, maintaining the original task of predicting molecular energy and forces while incorporating our method as an auxiliary task. In the context of coordinate denoising for equilibrium molecules, DP-TorchMD-NET and Frad have achieved notable results within the TorchMD framework by employing score-matching on the energy gradient [10, 29, 44]. In contrast, our denoising method does not require pre-training and can be effectively applied to non-equilibrium molecules. Our experimental setup is designed to align with these prior works, facilitating subsequent comparisons.

5.1. MD17 Dataset

5.1.1. Dataset

The MD17 dataset comprises molecular dynamics simulations of eight small organic molecules [8, 34], encompassing between 150,000 and nearly 1 million conformations, along with their corresponding total energy and force labels. We select the challenging task of force prediction as our downstream objective. For training and validation, we utilize 950 and 50 distinct conformations, respectively, with the remaining conformations allocated to the test set. Model performance is evaluated using the mean absolute error (MAE).

5.1.2. Results

Our baselines include both 3D pre-training approaches as well as classical supervised models. We list the force prediction results of SchNet [34], DimeNet [15], PaiNN [35], TorchMD-NET [39], SE(3)-DDM [26], Frad [10] and SliDe [29]. We particularly focus on the Frad method, as it introduces a coordinate denoising approach for equilibrium molecules, thereby achieving impressive results. The backbone network of this method is TorchMD-NET, and we adopt its experimental setup for our experiments, incorporating our auxiliary task to enhance force field learning capabilities. We compare these models with our proposed approach, with the results summarized in Table 1.

As shown in Table 1, our NDeM-augmented model achieves state-of-the-art or competitive performance on all eight molecules. Notably, it secures the best results on 5 out of 8 molecules and is the only method besides SliDe to con-

Table 3. Performance comparison on the S2EF task of the OC20 dataset. We report Mean Absolute Error (MAE) for energy and forces on the validation and test sets. Lower values are better.

Model	S2EF validation		S2EF test	
	Energy MAE (meV) ↓	Force MAE (meV/Å) ↓	Energy MAE (meV) ↓	Force MAE (meV/Å) ↓
GemNet-OC-L-E	239	22.1	230	21.0
GemNet-OC-L-F	252	20.0	241	19.0
SCN L=6 K=16 (4-tap 2-band)	-	-	228	17.8
SCN L=8 K=20	-	-	237	17.2
eSCN L=6 K=20	243	17.1	236	16.2
EquiformerV2 ($\lambda_E = 4$, 31M)	232	16.3	228	15.5
EquiformerV2 ($\lambda_E = 2$, 153M)	230	14.6	227	13.8
EquiformerV2 ($\lambda_E = 4$, 153M)	227	15.0	219	14.2
EquiformerV2 + DeNS ($\lambda_E = 4$, 160M)	221	14.2	**216**	13.4
EquiformerV2 + NDeM ($\lambda_E = 2$, 157M)	229	**13.8**	227	**12.9**
EquiformerV2 + NDeM ($\lambda_E = 4$, 157M)	**219**	14.7	218	13.8

sistently place in the top-tier, demonstrating its robustness on this challenging non-equilibrium dataset.

5.2. QM9 Dataset

5.2.1. Dataset

The QM9 dataset [31, 32] is a quantum chemistry collection that provides a single equilibrium conformation along with 12 labels for geometric, energetic, electronic, and thermodynamic properties, computed via density functional theory (DFT), for 134,000 stable small organic molecules composed of CHONF atoms. The dataset is partitioned following a standard configuration: 110,000 samples for the training set, 10,000 samples for the validation set, and the remaining 10,831 samples for the test set [11, 14, 33]. The performance across the 12 properties is evaluated using the mean absolute error (MAE, where lower values indicate better performance).

5.2.2. Results

Our baselines include both 3D pre-training approaches as well as classical supervised models. We list the results of SchNet [34], DimeNet++ [21], PaiNN [35], SphereNet [27], TorchMD-NET [39], Transformer-M [28], SE(3)-DDM [26], 3D-EMGP [19], Frad [10] and SliDe [29]. We compare these models with our proposed approach, with the results summarized in Table 2. The experimental findings demonstrate that, even on the QM9 dataset, which predominantly comprises equilibrium data, our training method performs robustly across multiple metrics. Notably, our approach achieves performance comparable to that of the Frad and SliDe methods, which are specifically tailored for equilibrium molecular data and have undergone pre-training. This suggests that our method can be generalized to all molecular data, irrespective of whether the molecules are in equilibrium or non-equilibrium states.

5.3. OC20 Dataset

5.3.1. Dataset

The expansive and diverse OC20 dataset encompasses approximately 1.2 million density functional theory (DFT) relaxation trajectories designated for training and evaluation [7]. Each structure within the OC20 dataset features an adsorbate molecule positioned on a catalyst surface, with the primary objective being the Structure to Energy and Forces (S2EF) task, which entails predicting the energy of the structure along with the per-atom forces. Models trained for the S2EF task are assessed based on the mean absolute error (MAE) for force predictions.

5.3.2. Results

We conducted experiments on the OC20 dataset, adopting EquiformerV2 [23] as the baseline model to ensure direct comparability with prior work. Our experimental standards follow those of the DeNS study [24], and we benchmark our results against methods such as GemNet-OC [16], SCN [48], and eSCN [30]. The results, presented in Table 3, demonstrate superior experimental performance.

5.4. Ablation

This subsection presents the ablation study on our auxiliary task, analyzes the contributions of our training method in practical applications, and discusses its impact on force field learning when integrated with the original task.

5.4.1. Auxiliary Task

The original force field learning task typically involves inputting raw molecular coordinates to predict energy and

Table 4. Ablation study on the auxiliary task coefficient λ_{NDeM} for force field learning on the Aspirin dataset (MD17).

λ_{NDeM}	Aspirin (MAE) ↓
0	0.2087
0.1	0.1688
1	0.1654
10	0.1671
100	0.1722

Table 5. Ablation study on the maximum standard deviation σ of multi-scale noise for coordinate denoising on the Aspirin dataset (MD17).

σ	Aspirin (MAE) ↓
0.005	0.1666
0.05	0.1654
0.1	0.1657
0.5	0.1887

molecular forces, with the experimental setup for the loss function referencing a ratio of $\lambda_E : \lambda_F = 1 : 100$, as force fields are generally more challenging to learn than energy. In this study, we investigate the optimal proportion of the auxiliary task that can enhance force field learning. The results of Aspirin Dataset in MD17, presented in Table 4, indicate that force field learning benefits when the auxiliary task coefficient λ_{NDeM} is set to an appropriate value.

5.4.2. Noise Scale

The physicochemical constraints for non-equilibrium molecules require the added noise to be sufficiently small to hold; excessive noise can disrupt the molecular structure. We adjust the maximum standard deviation of the multi-scale noise, with the experimental results presented in Table 5. The findings indicate that an appropriate noise scale facilitates the coordinate denoising process, consequently enhancing the model's generalization capability.

5.4.3. Compatibility with Pre-training

While NDeM enhances performance without requiring expensive pre-training, it is fully compatible with existing pre-trained checkpoints. We investigate this by initializing our model with weights from state-of-the-art methods, Frad [10] and SliDe [29], on the Aspirin dataset.

As shown in Table 6, incorporating NDeM during fine-tuning consistently outperforms standard continued training across different pre-trained baselines. These results indicate that NDeM effectively leverages pre-learned representations to further refine potential energy surface learning through its physically-grounded auxiliary task.

Table 6. Force prediction MAE on Aspirin (MD17) demonstrating the compatibility of NDeM with pre-trained checkpoints. PT denotes standard fine-tuning from the respective pre-trained weights.

Method	Aspirin (MAE) ↓
Frad PT [10]	0.2088
Frad + NDeM (Ours)	**0.1774**
SliDe PT [29]	0.1698
SliDe + NDeM (Ours)	**0.1648**
NDeM (From Scratch)	0.1654

5.4.4. Noisy Force Head

As illustrated in Figure 2, while the primary task follows a conventional equivariant GNN architecture to predict molecular energy and forces, a key design choice is the use of a separate *Noisy Force Head* for the auxiliary task. A pertinent question is why the auxiliary task, upon receiving noise-perturbed coordinates, does not simply reuse the primary force head given their identical output dimensions.

Directly outputting noisy forces through the primary head degrades original force prediction performance, whereas a dedicated auxiliary head ensures high learning precision. This requirement stems from the significant divergence between perturbed and unperturbed force directions in non-equilibrium molecules (Figure 1). Representing these physically distinct quantities with a single equivariant head induces optimization interference within the shared backbone, thereby compromising accuracy.

Furthermore, the new output head leverages latent representations of the original molecular forces, capturing a portion of the structural information. This allows the noisy force to be better aligned with the original molecular force, ultimately improving the model's capability in force field learning.

6. Conclusion

In this work, we addressed the limitations of traditional coordinate denoising for non-equilibrium structures by introducing Node Denoising on non-Equilibrium Molecules (NDeM). Leveraging a finite-difference approximation, NDeM decouples inherent forces from artificial noise, establishing a valid denoising objective for arbitrary conformations. Unlike pre-training approaches, NDeM functions as an efficient, architecture-agnostic auxiliary task seamlessly integrable into supervised pipelines. Extensive evaluations on standard benchmarks demonstrate that NDeM consistently enhances baseline performance, validating its robustness across both equilibrium and non-equilibrium regimes. Future work may explore higher-order Hessian approximations and applications to complex systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62495090, 62495094 and 62276127), Brain Science and Brain-like Intelligence Technology - National Science and Technology Major Project (Grant No. 2021ZD0201300), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118).

References

[1] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Enhancing drug-drug interaction extraction from texts by molecular structure information. *arXiv preprint arXiv:1805.05593*, 2018. 1

[2] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022. 1

[3] Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995. 2

[4] Ludwig Boltzmann. Studien uber das gleichgewicht der lebenden kraft. *Wissenschafiliche Abhandlungen*, 1:49–96, 1868. 2

[5] Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e (3) equivariant message passing. *arXiv preprint arXiv:2110.02905*, 2021. 2

[6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[7] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021. 7

[8] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017. 1, 6

[9] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[10] Shikun Feng, Yuyan Ni, Yanyan Lan, Zhi-Ming Ma, and Wei-Ying Ma. Fractional denoising for 3d molecular pre-training. In *International Conference on Machine Learning*, pages 9938–9961. PMLR, 2023. 1, 2, 6, 7, 8

[11] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pages 3165–3176. PMLR, 2020. 7

[12] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020. 2

[13] Zhangyang Gao, Cheng Tan, Jun Xia, and Stan Z Li. Cosupervised pre-training of pocket and ligand. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 405–421. Springer, 2023. 1

[14] Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020. 7

[15] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020. 6

[16] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C Lawrence Zitnick, and Abhishek Das. Gemnet-oc: developing graph neural networks for large and diverse molecular simulation datasets. *arXiv preprint arXiv:2204.02782*, 2022. 7

[17] Jonathan Godwin, Michael Schaarschmidt, Alexander Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Veličković, James Kirkpatrick, and Peter Battaglia. Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971*, 2021. 2

[18] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 2

[19] Rui Jiao, Jiaqi Han, Wenbing Huang, Yu Rong, and Yang Liu. Energy-motivated equivariant pretraining for 3d molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8096–8104, 2023. 1, 7

[20] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *Advances in neural information processing systems*, 35:24240–24253, 2022. 1

[21] Johannes Klicpera, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020. 7

[22] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022. 2

[23] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023. 2, 7

[24] Yi-Lun Liao, Tess Smidt, Muhammed Shuaibi, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *arXiv preprint arXiv:2403.09549*, 2024. 2, 5, 7

[25] Xinyuan Lin, Chi Xu, Zhaoping Xiong, Xinfeng Zhang, Ningxi Ni, Bolin Ni, Jianlong Chang, Ruiqing Pan, Zidong Wang, Fan Yu, et al. Pangu drug model: learn a molecule like a human. *Biorxiv*, pages 2022–03, 2022. 1

[26] Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with se (3)-invariant denoising distance matching. *arXiv preprint arXiv:2206.13602*, 2022. 6, 7

[27] Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2022. 7

[28] Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. One transformer can understand both 2d & 3d molecular data. *arXiv preprint arXiv:2210.01765*, 2022. 7

[29] Yuyan Ni, Shikun Feng, Weiying Ma, Zhiming Ma, and Yanyan Lan. Sliced denoising: A physics-informed molecular pre-training method. In *12th International Conference on Learning Representations (ICLR 2024)*. International Conference on Learning Representations, ICLR, 2024. 6, 7, 8

[30] Saro Passaro and C Lawrence Zitnick. Reducing so (3) convolutions to so (2) for efficient equivariant gnns. In *International conference on machine learning*, pages 27420–27438. PMLR, 2023. 2, 7

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

[32] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. 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. 7

[33] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021. 7

[34] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017. 6, 7

[35] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021. 6, 7

[36] Jocelyn Sietsma and Robert JF Dow. Creating artificial neural networks that generalize. *Neural networks*, 4(1):67–79, 1991. 2

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[38] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pages 20479–20502. PMLR, 2022. 1

[39] Philipp Thölke and Gianni De Fabritiis. Torchmd-net: equivariant transformers for neural network based molecular potentials. *arXiv preprint arXiv:2202.02541*, 2022. 6, 7

[40] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 2

[41] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 2, 3

[42] Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *Journal of Chemical Information and Modeling*, 62 (11):2713–2725, 2022. 1

[43] Dongyu Xue, Han Zhang, Dongling Xiao, Yukang Gong, Guohui Chuai, Yu Sun, Hao Tian, Hua Wu, Yukun Li, and Qi Liu. X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis. *bioRxiv*, pages 2020–12, 2020. 1

[44] Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133*, 2022. 1, 2, 3, 6

[45] Xiao-Chen Zhang, Cheng-Kun Wu, Zhi-Jiang Yang, Zhen-Xing Wu, Jia-Cai Yi, Chang-Yu Hsieh, Ting-Jun Hou, and Dong-Sheng Cao. Mg-bert: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings in bioinformatics*, 22(6):bbab152, 2021. 1

[46] Liangzhen Zheng, Jingrong Fan, and Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega*, 4(14):15956–15965, 2019. 1

[47] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. 2023. 1

[48] Larry Zitnick, Abhishek Das, Adeesh Kolluru, Janice Lan, Muhammed Shuaibi, Anuroop Sriram, Zachary Ulissi, and Brandon Wood. Spherical channels for modeling atomic interactions. *Advances in Neural Information Processing Systems*, 35:8054–8067, 2022. 7