Coordinate Denoising for Non-Equilibrium Molecular Representation Learning

Supplementary Material

7. Proof of Equivalence between Score Matching and Conditional Score Matching

The following proposition establishes the equivalence between score matching and conditional score matching, which is a key theoretical foundation for denoising approach [44].

Proposition 1. *The equivalence between score matching and conditional score matching. The two minimization objectives below are equivalent, i.e.,* $J_1(\theta) \approx J_2(\theta)$,

$$J_1(\theta) = E_{p(\tilde{x})}[\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x})\|^2] \tag{19}$$

$$J_2(\theta) = E_{p(\tilde{x}|x)p(x)}[\|GNN_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}|x)\|^2] \tag{20}$$

Proof. We first expand the square term in $J_1(\theta)$ and observe:

$$\begin{aligned} J_1(\theta) =& E_{p(\tilde{x})}[\|GNN_\theta(\tilde{x})\|^2] \\ & -2E_{p(\tilde{x})}[(GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \log p(\tilde{x}))] + T_3, \end{aligned} \tag{21}$$

$$\begin{aligned} J_2(\theta) =& E_{p(\tilde{x}|x)p(x)}[\|GNN_\theta(\tilde{x})\|^2] \\ & -2E_{p(\tilde{x})p(x|\tilde{x})}[(GNN_\theta(\tilde{x}), \nabla_{\tilde{x}} \log p(\tilde{x}|x))] + T_4, \end{aligned} \tag{22}$$

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 equal:

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

which is equal to the second term in $J_2(\theta)$, completing the proof. □

8. Comparison with Prior Denoising Approaches

Coordinate denoising represents an innovative approach in molecular force field prediction, leveraging the three-dimensional coordinate information of molecules. By incorporating physicochemical theoretical constraints, it achieves superior outcomes compared to methods relying solely on property prediction.

Godwin et al. [18] were the pioneers in proposing the addition of noise to 3D coordinates, subsequently employing denoising as an auxiliary task. This auxiliary task is trained concurrently with the primary task, eliminating the need for an additional large dataset. However, their method is constrained to known equilibrium structures, limiting its applicability to datasets such as QM9 and OC20 IS2RE [7, 33, 34], and rendering it unsuitable for force prediction tasks, such as those on the MD17 dataset [8, 9, 36]. In contrast, our study introduces force encoding to extend their framework, enabling its application to non-equilibrium structures. These structures constitute a significantly larger dataset compared to equilibrium ones, and our approach demonstrates enhanced performance on force field prediction datasets.

Zaidi et al. [48] adopted the denoising strategy proposed by Godwin et al. as a pre-training technique, necessitating an additional large dataset of unlabeled equilibrium structures for pre-training. Conversely, both Godwin et al. and our work integrate denoising directly with the primary task, avoiding the use of any supplementary unlabeled data.

Feng et al. [11] built upon the pre-training denoising methodology of Zaidi et al., introducing a distinct noise addition strategy. Specifically, they differentiate noise into dihedral angle noise and coordinate noise, focusing solely on predicting the latter. However, adding noise to dihedral angles requires tools like RDKit to identify rotatable bonds, which restricts its applicability to datasets such as OC20.

Furthermore, the aforementioned studies initially pre-train their models on the PCQM4Mv2 dataset [30] before fine-tuning on datasets like MD17. It should be noted that their experimental setup differs from ours, as we do not employ any dataset for pre-training purposes.

9. Experimental Setup

The hyperparameter settings for our experiments are presented in Table 7 and Table 8. For the components of the original task, excluding the auxiliary task, we largely align our configurations with those of the Frad method to facilitate a fair comparison.

Table 7. Hyperparameters for fine-tuning on the MD17 dataset.

Parameter	Setting
Train/Val/Test Splitting	950/50/remaining data
Batch size	8
Optimizer	AdamW
Warm up steps	1000
Max Learning rate	0.001
Learning rate decay policy	ReduceLROnPlateau 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	30
Energy weight	1
Denoising weight	1
Coordinate noise scale	0.05

Table 8. Hyperparameters for fine-tuning on the QM9 dataset.

Parameter	Setting
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
Network structure	TorchMD-NET
Head number	8
Layer number	8
RBF number	64
Activation function	SiLU
Embedding dimension	256
Head	256
AtomRef	1
Label weight	1
Noisy Nodes denoise weight	0.5
Coordinate noise scale	0.01

10. Denoising and Score Matching: A Theoretical Perspective

The application of noise to enhance the generalization capabilities of neural networks has been a well-established concept in the literature [3]. Denoising autoencoders [45], for instance, have been developed as a robust strategy to extract meaningful and resilient representations by interpreting the denoising process as a mechanism to capture the underlying data manifold. In the context of graph neural networks (GNNs) [18], the incorporation of noisy inputs during training has been shown to yield performance improvements. Notably, the Noisy Nodes approach employs denoising as an auxiliary loss to mitigate the issue of over-smoothing, thereby facilitating more accurate predictions of molecular properties.

Score matching, on the other hand, serves as an energy-based generative modeling technique designed to perform maximum likelihood estimation for unnormalized probability density models, particularly when the partition function is computationally intractable. A significant theoretical connection exists between denoising and score matching, especially when the noise follows a standard Gaussian distribution [44]. This relationship has been effectively leveraged in generative modeling and energy-based molecular modeling, particularly for learning force fields [40]. Despite sharing a common theoretical foundation, generative models and force field learning diverge in their practical assumptions and objectives, reflecting their distinct goals in application.