GeoTMI: Predicting Quantum Chemical Property with Easy-to-Obtain Geometry via Positional Denoising

Hyeonsu Kim[*]
Department of Chemistry
KAIST
Daejeon, South Korea

Jeheon Woo[*]
Department of Chemistry
KAIST
Daejeon, South Korea

Seonghwan Kim[*]
Department of Chemistry
KAIST
Daejeon, South Korea

Seokhyun Moon[*]
Department of Chemistry
KAIST
Daejeon, South Korea

Jun Hyeong Kim[*]
Department of Chemistry
KAIST
Daejeon, South Korea

Woo Youn Kim[†]
Department of Chemistry
KAIST
Daejeon, South Korea

Abstract

As quantum chemical properties have a dependence on their geometries, graph neural networks (GNNs) using 3D geometric information have achieved high prediction accuracy in many tasks. However, they often require 3D geometries obtained from high-level quantum mechanical calculations, which are practically infeasible, limiting their applicability to real-world problems. To tackle this, we propose a new training framework, GeoTMI, that employs denoising process to predict properties accurately using easy-to-obtain geometries (corrupted versions of correct geometries, such as those obtained from low-level calculations). Our starting point was the idea that the correct geometry is the best description of the target property. Hence, to incorporate information of the correct, GeoTMI aims to maximize mutual information between three variables: the correct and the corrupted geometries and the property. GeoTMI also explicitly updates the corrupted input to approach the correct geometry as it passes through the GNN layers, contributing to more effective denoising. We investigated the performance of the proposed method using 3D GNNs for three prediction tasks: molecular properties, a chemical reaction property, and relaxed energy in a heterogeneous catalytic system. Our results showed consistent improvements in accuracy across various tasks, demonstrating the effectiveness and robustness of GeoTMI.

1 Introduction

Neural networks have been actively applied to various fields of molecular and quantum chemistry [1, 2, 3, 4]. Several input representations, such as the SMILES string and graph-based representations, are employed to predict quantum chemical properties [5, 6]. In particular, graph neural networks (GNNs), which operate on molecular graphs by updating the representation of each atom via message-passing based on chemical bonds, have achieved great success in many molecular property prediction tasks [7, 8, 9, 10, 11, 12].

However, as many quantum chemical properties depend on molecular geometries, typical GNNs without 3D geometric information have limitations in their accuracy. In this respect, GNNs utilizing

[*]Equal contributors.
[†]Corresponding author: wooyoun@kaist.ac.kr

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3D information have recently achieved state-of-the-art accuracy [7, 13, 14, 15, 16, 17, 18]. Despite of their impressive accuracy, the usage of the 3D input geometry is often infeasible in real-world applications, limiting the 3D GNNs' applicability [19, 20, 21, 22, 23]. Therefore, it is natural to train machine learning models to make predictions with relatively easy-to-obtain geometries. Several studies have investigated the use of easy-to-obtain geometry as input, and it has been empirically confirmed that such geometry can be leveraged to accurately predict target properties [19, 22, 23]. Yet, theoretical basis for fully exploiting such easy-to-obtain geometries to predict accurate target properties remains to be established.

This study proposes a novel training framework, namely "**Geo**metric denoising for **T**hree-term **M**utual **I**nformation maximization (GeoTMI)", which employs a denoising process to accurately predict quantum chemical properties using easy-to-obtain geometries. Throughout this paper, we denote the correct geometry as X, the easy-to-obtain geometry (regarded as the corrupted version of X) as $\tilde{X}$, and the target property as Y. Various previous studies have been conducted on denoising approaches, such as a denoising autoencoder (DAE) [24, 25, 26, 27, 28, 29]. When it comes to predicting quantum chemical properties, the predominant focus of denoising techniques has been on improving the prediction accuracy starting from X. GeoTMI, however, aims at improving the prediction accuracy starting from $\tilde{X}$. GeoTMI also explicitly updates the input geometry of $\tilde{X}$ to approach X as it passes through the GNN layers, thereby contributing to more effective denoising. Furthermore, GeoTMI incorporates an auxiliary objective that predicts Y from X, allowing it to capture the task-relevant information and ultimately maximize the mutual information (MI) between the three terms of X, $\tilde{X}$, and Y. The theoretical derivations in this study provide further support for this approach.

GeoTMI offers the advantage of being model-agnostic and easy to integrate into existing GNN architectures. Thus, in this study, we aimed to validate the effectiveness of GeoTMI on different GNN architectures and for various prediction tasks (the nine other molecular properties of the QM9 [30], a chemical reaction property of Grambow's dataset [31], and relaxed energy in a heterogeneous catalytic system of the Open Catalyst 2020 (OC20) dataset [23]). We evaluated the performance of GeoTMI by comparing it to baselines trained only with $\tilde{X}$ and Y using multiple 3D GNNs. GeoTMI showed consistent accuracy improvements for all the target properties tested. In particular, in our experiment on the IS2RE task of the OC20, GeoTMI achieved greater performance improvements than another denoising method, Noise Nodes [27], demonstrating the superiority of GeoTMI. Overall, our findings demonstrate that GeoTMI can make accurate and robust predictions with easy-to-obtain geometries. Code is available on Github.

2 Related Works

2.1 Predicting high-level properties from easy-to-obtain geometry

Recently, several deep learning approaches have aimed to predict high-level properties from an easy-to-obtain geometry for accurate yet fast predictions in real-world applications. For instance, Molecule3D benchmark [22] aims to improve the applicability of existing 3D models by developing machine learning models that predict 3D geometry. These models predicted 3D geometry using 2D graph information that can be easily obtained and were evaluated using ETKDG [32] from RDKit [33] as a baseline.

There have been attempts to predict high-level properties, starting with a geometry that can be quickly obtained by conventional methods, rather than machine learning methods. Lu et al. [19] adopted Merck molecular force field (MMFF) [34] geometries as the starting point, to predict density functional theory (DFT) [35] properties of the molecules in the QM9 dataset. In chemical reactions, Spiekermann et al. [36] exploited reactant and product geometries to assess reaction barrier height rather than reactant and transition state (TS) geometries; because obtaining the TS geometry is computationally challenging. In addition, Chanussot et al. [23] proposed the Open Catalyst challenge. In this challenge, the IS2RE task uses initial structures (IS) for geometry optimization to predict the relaxed energies (RE) of the corresponding relaxed structures (RS). In this case, the IS and the RE can be mapped into easy-to-obtain geometries and high-level properties, respectively. Various approaches have been proposed to address this challenge [16, 27, 37].

GeoTMI shares the same goal as these previous works. However, it is important to emphasize that we propose a training framework based on a theoretical basis that possesses the capacity to be applicable to various tasks, rather than being limited to a specific task.

2.2 Denoising approaches in GNN

Denoising is a commonly used approach for representation learning by recovering correct data from corrupted data. Previous studies have shown that models can learn desirable representations by mapping from a corrupted data manifold to a correct data manifold. Traditional denoising auto-encoders (DAEs) employed a straightforward procedure of recovering a correct data from corrupted data thus maximizes the MI between correct data and its representation [24, 25]. Recently, several studies in GNNs have adopted denoising strategies for representation learning and robustness during training [28, 29, 38, 39, 40]. For instance, Noisy Nodes [27], which primary aim is addressing oversmoothing in GNNs, used denoising noisy node information as an auxiliary task, resulting in improved performance in property prediction. Additionally, LaGraph [41] leveraged predictive self-supervised learning of GNNs to predict the intractable latent graph that represents semantic information of an observed graph, by introducing a surrogate loss originated from image representation learning [42]. While typical denoising approaches focus on learning representations of expensive X, GeoTMI aims to learn higher-quality representations for $\tilde{X}$, lying on a geometrically corrupted data manifold, to predict Y. For this purpose, GeoTMI adopts the maximization of the three-term MI between X, $\tilde{X}$, and Y with theoretical basis.

2.3 Invariant 3D GNNs for quantum chemical properties

In the field of chemistry, GNNs utilizing 3D geometric information have shown promising performance in predicting quantum chemical or systematic properties [7, 13, 14, 15, 37, 43, 44, 45]. Since target physical quantities, such as an energy, are invariant to alignments of a molecule, 3D GNN models utilize roto-translational invariant 3D information as their inputs [46, 47]. As a representative example, a distance matrix guarantees the invariance because the roto-translational transformation does not vary distances. SchNet [7, 43] and EGNN [15] are proper examples of utilizing the distance matrix. The former exploits the radial basis function based on the distance matrix, while the latter uses distance information directly on the GNN message-passing scheme. In DimeNet++ [44], along with the distance matrix, bond angles are also available as invariant 3D information. In addition, ComENet [45] and SphereNet [14] introduced dihedral angles in addition to the distance and bond angle information. Recently, several approaches such as Equiformer [16] explicitly considered irreducible representations to construct roto-translational equivariant neural networks [48, 49]. Our evaluation showed that GeoTMI is model-agnostic, hence can be easily applied to various 3D GNNs.

3 Method

In this section, we describe the overall framework of our proposed GeoTMI with theoretical background. First, in Section 3.1, we introduce the problem setting, i.e., the physical relationships required to predict a property from a corrupted geometry. Then, in Section 3.2, we introduce our training objective for three-term MI, which differs from the objective of typical supervised learning. Since MI itself is intractable, we derive a tractable loss for the training objective in Section 3.3. Finally, in Section 3.4, we illustrate the practical application of GeoTMI framework in the training and inference processes.

3.1 Problem setup

We first introduce physical relationship between our data: corrupted geometry, $\tilde{X}$, correct geometry, X, and quantum chemical property, Y. Our training data, $\mathcal{D}$, consist of observed samples $(\tilde{x}, x, y)$ from the triplet of three random variables $(\tilde{X}, X, Y) \sim q(\tilde{X}, X, Y)$. We assume that these three variables are interlinked through a Markov chain $\tilde{X} \rightarrow X \rightarrow Y$. In our problem setting, $\tilde{Z}$ and Z denote representations of $\tilde{X}$ and X, respectively, whose probability distributions are parameterized by θ, $p_\theta(\tilde{Z}, Z, Y)$.

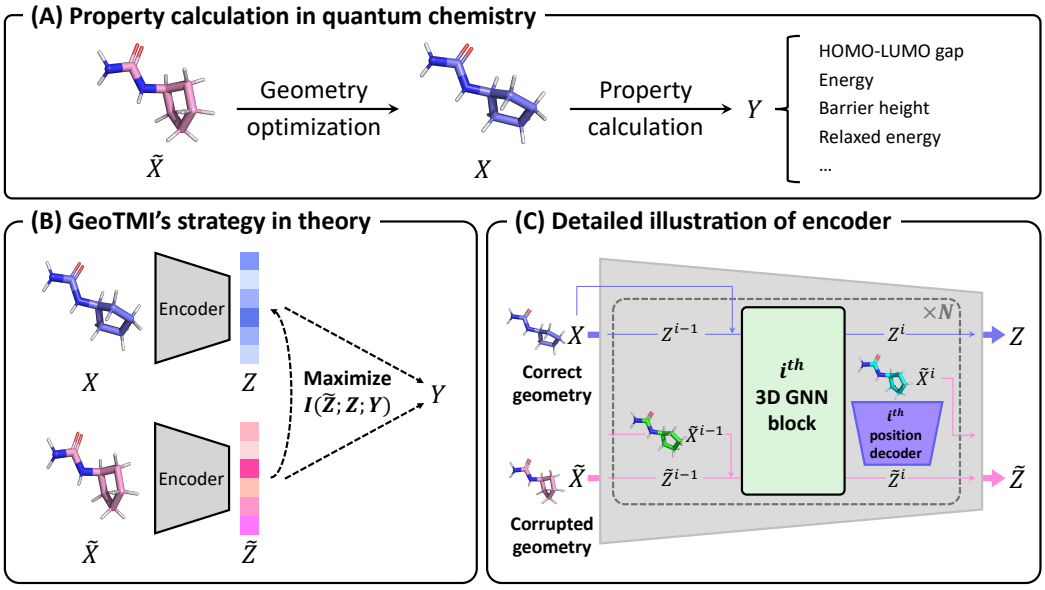

Figure 1: (A) Physical relationship between $\tilde{X}$, X, and Y in property calculation in quantum chemistry. (B) Schematic illustration of GeoTMI's strategy in theory, where objective is maximizing three-term MI, $I(\tilde{Z}; Z; Y)$. (C) Detailed illustration of encoder architecture in practical strategy of GeoTMI. Training process employs both blue and pink lines, while inference process utilizes only pink line. All molecular geometries were plotted using PyMOL [50].

Within the standard computational chemistry process, Y is obtained from the correct geometry X, which is acquired through geometric optimization of $\tilde{X}$ as shown in Figure 1(A). This process naturally gives rise to the Markov chain assumption, which suggests that X encapsulates all the essential information for Y. We can establish two assumptions employing the underlying physical relationship between the variables as an inductive bias. First, there exists a higher quality of information pertaining to Y within X compared to $\tilde{X}$, or, more precisely, the MI between X and Y that is equal to or greater than the MI between $\tilde{X}$ and Y. This naturally follows from the property of conditional independence between non-adjacent states in a Markov chain. Second, the data distribution of Y is solely dependent on X, irrespective of the presence of $\tilde{X}$, implying that $\tilde{X}$ and Y are conditional independent given X.

The goal of GeoTMI is to obtain a proper representation $\tilde{Z}$ in predicting Y, by aligning it into Z that contains more enriching information for Y. GeoTMI differs to self-supervised learning by emphasizing a specialized representation that is tailored to the target property. Also, acquisition of physical relationship between the variables as inductive bias leads to a higher quality representation than focusing on predicting Y using $\tilde{X}$ alone.

3.2 Training objective

We propose a training framework for learning a proper representation for predicting Y from $\tilde{X}$, which can be done by maximizing the MI between the variables, $I_\theta(\tilde{Z}; Y)$. This is somewhat similar to the objective of the general supervised learning which predicts Y from $\tilde{X}$. However, the training only with $\tilde{X}$ to predict Y could be erroneous because there is no guarantee a model utilize the proper information resided in both $\tilde{X}$ and X.

If we introduce Z, one can express $I_\theta(\tilde{Z}; Y)$ as following:

$$I_\theta(\tilde{Z}; Y) = I_\theta(\tilde{Z}; Y|Z) + I_\theta(\tilde{Z}; Z; Y). \tag{1}$$

In Equation (1), the conditional MI $I_\theta(\tilde{Z}; Y|Z)$ implies undesirable information of $\tilde{Z}$ in predicting Y that is not relevant to Z. This is a direct counterpart to the physical inductive bias in the previous

section, where X is sufficient information for the prediction of Y, and thus should be minimized to zero in the optimal case (see Appendix A.1). Maximizing $I_\theta(\tilde{Z}; Y)$ while maintaining the zero inductive bias is ideal, but non-trivial and challenging. To address this, we propose a straightforward solution by introducing the inductive bias as a regularization term, which reformulates our initial objective into maximization of the three-term MI,

$$I_\theta(\tilde{Z}; Z; Y) = I_\theta(\tilde{Z}; Y) - I_\theta(\tilde{Z}; Y|Z).$$

3.3 Tractable loss derivation

In general, MI is not tractable, and accurately estimating it is another challenging task. Thus, we have derived a tractable lower bound of the three-term MI, allowing us to practically maximize it (see Figure 1(B)).

Proposition 3.1. *$I(\tilde{Z}; Z; Y) \geq \mathrm{LB} + H(Y)$ for any triplets of random variables $(\tilde{Z}, Z, Y)$, where $\mathrm{LB} = -H(Y|Z) - \frac{1}{2}H(Y|\tilde{Z}) - \frac{1}{2}H(Z|\tilde{Z})$.*

Since $H(Y)$ is constant term in respect of model parameters, we have the three distincts optimization targets: (1) property from corrupted representation $H(Y|\tilde{Z})$, (2) property from correct representation $H(Y|Z)$, and (3) reconstruction to correct representation $H(Z|\tilde{Z})$.

The conditional entropy term related to the property is estimated by a parameterized distribution p_{π_1} based on the positiveness of KL divergence:

$$\begin{aligned} -H(Y|Z) - H(Y|\tilde{Z}) &= \mathbb{E}_{p_\theta(\tilde{Z},Z,Y)}\left[\log p_\theta(Y|\tilde{Z}) + \log p_\theta(Y|Z)\right] \\ &\geq \mathbb{E}_{p_\theta(\tilde{Z},Z,Y)}\left[\log p_{\pi_1}(Y|\tilde{Z}) + \log p_{\pi_1}(Y|Z)\right] \\ &\sim -\sum \left(\mathcal{L}\left(y, h_{\pi_1}(\tilde{z})\right) + \mathcal{L}\left(y, h_{\pi_1}(z)\right)\right). \end{aligned}$$

Here, we introduce property predictor h_{π_1} which is parameterized by π_1. Similarly, the other term is estimated by a parameterized distribution p_{π_2},

$$\begin{aligned} -H(Z|\tilde{Z}) &= \mathbb{E}_{p_\theta(\tilde{Z},Z,Y)}\left[\log p_\theta(Z|\tilde{Z})\right] \\ &\geq \mathbb{E}_{p_\theta(\tilde{Z},Z,Y)}\left[\log p_{\pi_2}(Z|\tilde{Z})\right] \\ &\sim -\sum \mathcal{L}(z, \hat{g}_{\pi_2}(\tilde{z})). \end{aligned}$$

Here, $\hat{g}_{\pi_2} : \tilde{Z} \to Z$ denotes a parametric decoder for information flows. Since Z is a parameterized variable which is not optimal in an initial training stage, the optimization could be unstable. If we assume the encoder $f_\theta : X \to Z$ is continuous bijective, we could introduce a surrogate loss of decoding $\tilde{Z}$ into X,

$$\mathbb{E}_{p_\theta(\tilde{Z},Z,Y)}\left[\mathcal{L}\left(f_\theta^{-1}(Z), f_\theta^{-1} \circ \hat{g}_{\pi_2}(\tilde{Z})\right)\right] = \mathbb{E}_{p_\theta(\tilde{Z},Z,Y)}\left[\mathcal{L}\left(X, g_{\pi_2}(\tilde{Z})\right)\right],$$

where $g_{\pi_2} = f_\theta^{-1} \circ \hat{g}_{\pi_2} : \tilde{Z} \to X$ denotes a decoder reconstructing X rather than Z. It still maximizes MI between $\tilde{Z}$ and Z, in that the continuous and bijective mapping does not change the MI. In summary, the training process is about finding optimal model parameters θ, π_1, and π_2 to minimize the following:

$$\mathbb{E}_{p_\theta(Z,\tilde{Z},Y)}\Bigg[\underbrace{\mathcal{L}\big(Y, h_{\pi_1}(\tilde{Z})\big)}_{\mathcal{L}_{\mathrm{y,corrupted}}} + \underbrace{\mathcal{L}\big(Y, h_{\pi_1}(Z)\big)}_{\mathcal{L}_{\mathrm{y,correct}}} + \underbrace{\mathcal{L}\big(X, g_{\pi_2}(\tilde{Z})\big)}_{\mathcal{L}_d}\Bigg].$$

The tractable loss function comprises the three terms: $\mathcal{L}_{\mathrm{y,corrupted}}$, $\mathcal{L}_{\mathrm{y,correct}}$, and $\mathcal{L}_d$. We refer to $\mathcal{L}_{\mathrm{y}}$ as the property prediction loss and $\mathcal{L}_d$ as the denoising loss. We chose the absolute error for the loss function $\mathcal{L}$. The proof of Proposition 3.1 and details of the denoising loss are described in Appendix A.

3.4 Overall framework

The proposing framework comprises the encoder, predictor, and decoder. The encoder maps molecular geometries to their representations, while the predictor estimates target properties, and the decoder restores the molecular geometries. The encoder design involves 3D GNN layers for both X and $\tilde{X}$, sharing model parameters. It is appropriate approach because X and $\tilde{X}$ belong to the same data modality. In addition, the encoder for $\tilde{X}$ includes explicit position update layers that are inspired by the geometry optimization process. The effect of intermediate geometries as input is studied in Appendix B.1. The practical model architecture including the encoder design is depicted in Figure 1(C).

In practice, an auxiliary loss is introduced as an add-on for the denoising loss to softly guide the position update toward X. We will refer to this as gradual denoising loss, which measures the difference between each updated geometry and the corresponding linearly interpolated target geometry. The details and ablation study of this are in Appendix B.2. We chose the same architecture of the position update layer for the decoder.

During training, $\tilde{X}$ and X are mapped to $\tilde{Z}$ and Z respectively. The property prediction loss is computed based on the results from $\tilde{Z}$ and Z, while the denoising loss involves reconstruction of X from $\tilde{Z}$. In inference process, only $\tilde{Z}$ encoded by $\tilde{X}$ is used for property prediction. It is noteworthy that GeoTMI introduces a novel representation learning approach that leverages X and Z for robust property prediction, and its effectiveness lies in not requiring X and Z during the inference process.

4 Experiments

We have tried to demonstrate the effectiveness of GeoTMI in providing a new solution to the infeasibility of high-level 3D geometry, rather than focusing on the performance of the state-of-the-art GNN architecture itself. Thus, in this section, we have focused on showing the applicability of GeoTMI to a variety of GNN architectures and its effectiveness in predicting properties in various areas of chemistry. The tasks and architectures tested were selected based on computational cost and memory efficiency, as well as model performance. All experiments were conducted using RTX 2080 Ti GPU with 12 GB of memory, RTX 3080 Ti GPU with 12 GB of memory, or RTX A4000 GPU with 16 GB of memory. GNN models were trained on a single GPU, except for those in the IS2RE task of OC20, where we used eight RTX A4000 GPUs.

4.1 Molecular property prediction

Predicting molecular properties is crucial to various fields in chemistry. The QM9 [30] is widely used benchmark dataset for molecular property prediction comprised of 134k molecule information; each molecule consists of at most nine heavy atoms (C, N, O, and F). Each data sample contains optimized geometry and more than 10 corresponding DFT properties.

This study focuses on the QM9_{M} [19] task which predicts DFT properties using the MMFF geometry. The QM9_{M} dataset originated from the QM9 differs only in the molecular geometry; each geometry herein has been obtained with additional MMFF optimization starting with the corresponding geometry in the QM9. Here, the MMFF geometry is regarded as a relatively easy-to-obtain geometry compared to the DFT geometry.

Training setup. This study employed the following three GNNs using distinct 3D information to demonstrate the effectiveness of GeoTMI: EGNN [15] (implementation follows [51]), SchNet [43], and DimeNet++ [44]. We appended the position update to the SchNet and DimeNet++ to ensure that the denoising process can be applied to them in the same manner in the coordinate update of the EGNN. To train the models, we considered the DFT geometry from the QM9 dataset as X, and the corresponding MMFF geometry from QM9_{M} dataset as $\tilde{X}$.

Molecular 2D graph information is similar to MMFF geometry information in that it is also more readily available than DFT geometry. There have been many attempts to predict accurate molecular properties from 2D graphs alone [17, 52, 53]. Recently, Luo et al. [17] developed the Transformer-M model, which can utilize both 2D graph and 3D geometry information in training to predict molecular properties with high accuracy using only 2D graphs. To compare the usefulness of 2D graphs

and MMFF geometries as easy-to-obtain inputs, we evaluated the prediction performance of the Transformer-M model on 2D graph inputs without pre-training. Note that the Transformer-M model reported their performance using X based on pre-training in the original paper.

For all tested models, we used 100,000, 18,000, and 13,000 molecular data for training, validation, and testing, respectively, as in previous work by Satorras et al. [15]. The detailed hyperparameters of each model are introduced in Appendix C.1.

Results on molecular property prediction. Table 1 shows the prediction accuracy according to input types and models. Results for SchNet and DimeNet++ are shown in Appendix B.3. GeoTMI achieved performance improvements across all properties and models. For example, GeoTMI resulted in accuracy improvements of 7.0∼27.1% for EGNN, as shown in Table 1. Meanwhile, Transformer-M trained using both 2D graphs and X resulted in accuracy improvements of -15∼21% compared to the same model trained using 2D graphs only. Despite the similar prediction performance of Transformer-M and EGNN based on X, it is noteworthy that for most properties, the Transformer-M models using 2D graphs for prediction were less accurate than the 3D GNNs tested. This result implies that while both the MMFF geometry and the molecular the 2D graph are easy-to-obtain inputs, the MMFF geometry contains more useful information for learning the relationship between molecules and their quantum chemical properties. Furthermore, we conducted additional experiments for three properties (μ, R^2, and U_0) using scaffold-based splitting, a methodology that offers a more realistic and demanding setting for evaluating out-of-distribution (OOD) generalization (see Appendix B.4). Once again, GeoTMI consistently improved its prediction performance, highlighting the robustness of GeoTMI.

Table 1: MAEs for QM9's properties. The best performance among the models that do not use X in the inference (Infer.) process is shown in bold. The values of Transformer-M using X were borrowed from Luo et al. [17]. The performance of GeoTMI integrated with SchNet and DimeNet++ is provided in Appendix B.3.

Methods	Input type (Train / Infer.)	U_0 (meV)	μ (D)	α (Bohr^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	GAP (meV)	R^2 (Bohr^2)	C_v ($\frac{cal}{mol \cdot K}$)	ZPVE (meV)
Transformer-M [17]	X/X	14.8	-	-	26.5	23.8	-	-	-	-
EGNN	X/X	12.9	0.0350	0.0759	31.2	26.6	51.1	0.130	0.0336	1.59
Transformer-M	2D / 2D	38.2	0.309	0.171	53.6	52.5	77.1	11.4	0.0669	4.79
Transformer-M	2D, X / 2D	43.9	0.245	0.160	48.7	46.3	68.4	10.3	0.0683	3.85
EGNN	$\tilde{X}/\tilde{X}$	17.4	0.133	0.125	38.4	34.4	58.0	5.60	0.0445	1.97
EGNN + GeoTMI	$X, \tilde{X}/\tilde{X}$	**14.5**	**0.100**	**0.105**	**35.7**	**31.2**	**53.2**	**4.08**	**0.0407**	**1.76**
Improvements by GeoTMI (%)		16.7	24.8	16.0	7.03	9.30	8.28	27.1	8.54	10.7

4.2 Reaction property prediction

A chemical reaction is a process in which reactant molecules convert to product molecules, passing through their TSs. Predicting properties related to the reaction is important for understanding the nature of the chemistry [31]. The barrier height, as one of the reaction properties, is defined as the energy difference between the geometry of the TS, X^{TS}, and the geometry of the reactant, X^R. Commonly, optimizing a X^{TS} utilizes both X^R and the geometry of the product X^P. However, this optimization process is typically resource-intensive, requiring approximately 10 times more computational resources than optimizing X^R or X^P individually [54]. Thus, predicting accurate barrier height without X^{TS} is necessary to reduce computational costs.

From this point of view, we focused on the task to predict DFT calculation-based barrier heights using X^R and X^P for the elementary reaction of the gas phase, as reported by Spiekermann et al. [36]. In contrast to most molecular properties, the property is a function of not just a single molecular geometry, but (X^R, X^{TS}), which can be interpreted as an optimized version of (X^R, X^P). Thus, we considered that $X := (X^R, X^{TS})$ and $\tilde{X} := (X^R, X^P)$ in this task.

We used two datasets, released by Grambow et al. [31], for comparison with the previous work. The first dataset consists of unimolecular reactions, namely CCSD(T)-UNI. The second dataset, B97-D3, has 16,365 reactions.

Training setup. Spiekermann et al. [55] proposed two models for predicting the barrier height using the 2D and 3D information of $\tilde{X}$, respectively. They used D-MPNN for 2D GNN and DimeNet++ for 3D GNN, which will be referred to as the 2D D-MPNN model and the 3D DimeReaction (DimeRxn), respectively. Here, the DimeRxn trained with $\tilde{X}$ showed lower performance than the 2D D-MPNN because the 3D GNNs were sensitive to the noise in the input geometry, as pointed out in another study [56]. Thus, our method, which removes the noise, can be useful for DimeRxn.

For the DimeRxn, we also adopted EGNN's coordinate update scheme as a decoder to predict correct geometries. We trained the D-MPNN model, and the DimeRxn models without and with GeoTMI for CCSD(T)-UNI and B97-D3 datasets. The used data split and augmentation were the same as in the previous work by Spiekermann et al. [55]. In particular, we note that scaffold splitting was used on the datasets to evaluate the OOD generalization ability of the model. The hyperparameters used are described in Appendix C.1.

Table 2: MAEs for predicted reaction barrier ($\mathrm{kcal/mol}$). The best performance among the models that do not use X in the inference (Infer.) process is shown in bold.

Methods	Input type (Train / Infer.)	Dataset	
		CCSD(T)-UNI	B97-D3
DimeRxn	X/X	2.38	1.92
D-MPNN	2D / 2D	4.59	4.91
DimeRxn	$\tilde{X}/\tilde{X}$	6.03	7.32
DimeRxn + GeoTMI	$X, \tilde{X}/\tilde{X}$	**3.90**	**4.17**
Improvements by GeoTMI (%)		35.3	43.0

Results on reaction property prediction. Table 2 shows the results of prediction accuracy according to input types and models. The DimeRxn trained with X has the best prediction performance for all methods, while DimeRxn trained with $\tilde{X}$ has the worst prediction performance. The result supports that DimeRxn is highly dependent on the quality of input geometry, as previously mentioned. Thus, as we expected, GeoTMI, which is developed for learning a proper representation for predicting Y from $\tilde{X}$, induced accuracy improvements of 35.4% and 43.0% than DimeRxn without GeoTMI in terms of MAE for the CCSD(T)-UNI and B97-D3 datasets, respectively. The results show that it outperforms the 2D D-MPNN model, again demonstrating the usefulness of the 3D easy-to-obtain geometry with GeoTMI, which is identified in the previous section.

4.3 IS2RE prediction

The OC20 dataset contains data consisting of the slab called a catalyst and molecules called adsorbates for each of the systems. There are more atoms and a wider variety of atom types compared to previously studied datasets. In detail, the dataset contains more than 460k pairs of IS, RS, and RE. We focus on the IS2RE task, which is to predict the RE using the IS. From the perspective of computational chemistry, the RS are obtained through costly quantum chemical calculations based on the IS. Thus, in this task, we considered IS as $\tilde{X}$ and RS as X.

Training setup. We adopted the Equiformer model [16] to evaluate the effectiveness of GeoTMI in the IS2RE task. The Equiformer model achieved state-of-the-art performance by using Noisy Nodes, where the IS2RS auxiliary task was integrated with the IS2RE task. We note that the hyperparameters used are the same as in the previous work, except for the number of transformer blocks to train each model, due to the limitation of our computational resources. We refer to the model trained only on the IS2RE task without Noisy Nodes as the baseline model, namely Equiformer*. We performed a comparative analysis of three training frameworks: (1) Equiformer*, (2) Equiformer* + Noisy Nodes, and (3) Equiformer* + GeoTMI. Thus, this evaluation with Equiformer can show the effectiveness of GeoTMI on the baseline model while allowing for comparison with Noisy Nodes.

To implement the Equiformer with GeoTMI, we followed much of the original Equiformer paper. First, we used the same Noisy Nodes data augmentation. Second, we used a similar node-level

auxiliary loss for the IS2RS task. The auxiliary loss predicts the node-level difference between target positions and noisy inputs, which corresponds to the denoising loss of $\mathcal{L}_d$. The different points of the "Equiformer* + GeoTMI" compared to the "Equiformer* + Noisy Nodes" are as follows. The noisy positions were explicitly updated by passing through GNN layers. The detailed objective here is to calculate the difference between the updated noisy positions and the linearly interpolated target positions at each GNN layer, which we refer to as the gradual denoising loss in our paper. In addition, we incorporated an auxiliary task that predicts the RE from the RS, denoted as $\mathcal{L}_{y,\text{correct}}$, which ultimately facilitates the training process of maximizing the three-term MI.

Table 3: Results on the OC20 IS2RE test set with different methods based on Equiformer architectures [16]. The Equiformer* denotes a model that reduces the number of transformer blocks from 18 to 4 while keeping other hyperparameters the same. The best performance among the Equiformer* models is shown in bold, and its improvement rate is shown in the last row.

	Energy MAE (eV) ↓					EwT (%) ↑				
Methods	ID	OOD Ads	OOD Cat	OOD Both	Average	ID	OOD Ads	OOD Cat	OOD Both	Average
Equiformer + Noisy Nodes [16]	0.417	0.548	0.425	0.474	0.466	7.71	3.70	7.15	4.07	5.66
Equiformer*	0.515	0.651	0.531	0.603	0.575	4.81	2.50	4.45	2.86	3.66
Equiformer* + Noisy Nodes	0.449	0.606	0.460	0.540	0.513	6.47	3.04	5.83	3.52	4.72
Equiformer* + GeoTMI	**0.425**	**0.583**	**0.440**	**0.521**	**0.492**	**7.60**	**3.86**	**6.97**	**4.03**	**5.62**
Improvement (%)	17.6	10.5	17.1	13.7	14.4	58.0	54.4	56.6	40.9	53.8

Results on IS2RE with Noisy Nodes. We have summarized the IS2RE results in Table 3. To evaluate each method, the MAE of the RE prediction using IS and the energy within a threshold (EwT), the percentage in which the MAE of the predicted energy is within 0.02 eV, are used. Both Noisy Nodes and GeoTMI show performance improvements over the baseline Equiformer*, but GeoTMI achieves better performance gains across all metrics. Despite the improvements, the prediction performance with GeoTMI is still lower than the original model with 18 transformer blocks in most cases in terms of MAE. However, the prediction performance is similar to the original model for EwT and even better for OOD Ads.

4.4 Ablation study

GeoTMI uses a combination of $\mathcal{L}_d$, $\mathcal{L}_{\text{y,correct}}$, and the position update to improve the accuracy of predicting quantum chemical properties using $\tilde{X}$. To verify an individual contribution of each component of GeoTMI, we conducted ablation studies. Table 4 shows that all strategies are individually meaningful to reduce prediction error regardless of the properties. In this experiment, it is noteworthy that training without either $\mathcal{L}_{\text{y,correct}}$ or $\mathcal{L}_d$ is no longer maximizing the lower bound of the three-term MI. The prediction performance of these models performs worse than trained models using GeoTMI except for C_v. The results imply that our proposed three-term MI maximization is key in prediction performance based on $\tilde{X}$. Additionally, the table shows that position update, introduced

Table 4: Ablation study for GeoTMI. BH and PU denote reaction barrier height and position update, respectively. Prediction accuracy is compared in terms of MAE. The most degraded results are underlined.

Dataset	Property	Unit	GeoTMI	w/o $\mathcal{L}_d$	w/o PU	w/o $\mathcal{L}_{\text{y,correct}}$
QM9 + QM9$_{\text{M}}$	U_0	meV	14.5	<u>21.0</u>	14.2	15.2
	R^2	Bohr^2	4.08	<u>6.43</u>	4.12	4.43
	C_v	$\text{cal/mol} \cdot \text{K}$	0.0407	<u>0.0503</u>	0.0410	0.0401
	μ	D	0.0997	<u>0.127</u>	0.111	0.110
CCSD(T)-UNI	BH	kcal/mol	3.90	4.41	<u>5.77</u>	4.27
B97-D3	BH	kcal/mol	4.17	4.49	<u>7.19</u>	4.45

by our intuition, is a key component for the BH prediction and helps increase overall prediction performance.

5 Conclusion and Limitations

In this study, we propose GeoTMI, a novel training framework designed to exploit easy-to-obtain geometry for accurate prediction of quantum chemical properties. The proposed framework is based on the Markov chain assumption and the theoretical basis that maximizes the mutual information (MI) between property, correct and corrupted geometries, mitigating the degradation in accuracy resulting from the use of the corrupted geometry. To achieve this, GeoTMI incorporates a denoising process to effectively address the inherent challenges associated with acquiring correct 3D geometry. In particular, we introduced the position update in the denoising process and gradual denoising loss to enhance the efficacy of the training process.

We have verified that GeoTMI consistently improves the prediction performance of 3D GNNs for three benchmark datasets. Nevertheless, there are several limitations in this work. First, GeoTMI addresses the inductive bias by incorporating a soft regularization approach instead of directly vanishing it to zero. Second, we could not perform an extensive optimal hyperparameter search due to a lack of computational resources. However, our consistent experimental results on various tasks showed the effectiveness and robustness of the GeoTMI. In this light, we envision that the GeoTMI becomes a new solution to solve the practical infeasibility of high-cost 3D geometry in many other chemistry fields.

6 Acknowledgement

This work was supported by the Korea Environmental Industry and Technology Institute (Grant No. RS202300219144), the Technology Innovation Program funded by the Ministry of Trade, Industry & Energy, MOTIE, Korea (Grant No. 20016007), and the Ministry of Science and ICT, Korea (Grant No. RS-2023-00257479).

References

[1] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1), November 2019. doi: 10.1038/s41467-019-12875-2. URL https://doi.org/10.1038/s41467-019-12875-2.

[2] Sergei Manzhos and Tucker Carrington. Neural network potential energy surfaces for small molecules and reactions. *Chemical Reviews*, 121(16):10187–10217, October 2020. doi: 10.1021/acs.chemrev.0c00665. URL https://doi.org/10.1021/acs.chemrev.0c00665.

[3] Pavlo O. Dral. Quantum chemistry in the age of machine learning. *The Journal of Physical Chemistry Letters*, 11(6):2336–2347, March 2020. doi: 10.1021/acs.jpclett.9b03664. URL https://doi.org/10.1021/acs.jpclett.9b03664.

[4] Sanggil Park, Herim Han, Hyungjun Kim, and Sunghwan Choi. Machine learning applications for chemical reactions. *Chemistry – An Asian Journal*, 17(14), May 2022. doi: 10.1002/asia.202200203. URL https://doi.org/10.1002/asia.202200203.

[5] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12 (1), September 2020. doi: 10.1186/s13321-020-00460-5. URL https://doi.org/10.1186/s13321-020-00460-5.

[6] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1), February 2021. doi: 10.1186/s13321-020-00479-8. URL https://doi.org/10.1186/s13321-020-00479-8.

[7] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24): 241722, June 2018. doi: 10.1063/1.5019779. URL https://doi.org/10.1063/1.5019779.

[8] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, July 2019. doi: 10.1021/acs.jcim.9b00237. URL https://doi.org/10.1021/acs.jcim.9b00237.

[9] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, August 2019. doi: 10.1021/acs.jmedchem.9b00959. URL https://doi.org/10.1021/acs.jmedchem.9b00959.

[10] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. doi: 10.1016/j.aiopen.2021.01.001. URL https://doi.org/10.1016/j.aiopen.2021.01.001.

[11] Seokhyun Moon, Wonho Zhung, Soojung Yang, Jaechang Lim, and Woo Youn Kim. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, 13(13):3661–3673, 2022. doi: 10.1039/d1sc06946b. URL https://doi.org/10.1039/d1sc06946b.

[12] Jun Hyeong Kim, Hyeonsu Kim, and Woo Youn Kim. Effect of molecular representation on deep learning performance for prediction of molecular electronic properties. *Bulletin of the Korean Chemical Society*, 43(5):645–649, March 2022. doi: 10.1002/bkcs.12516. URL https://doi.org/10.1002/bkcs.12516.

[13] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs, 2020. URL https://arxiv.org/abs/2003.03123.

[14] Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3D graph networks, 2021. URL https://arxiv.org/abs/2102.05013.

[15] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

[16] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3D atomistic graphs, 2022. URL https://arxiv.org/abs/2206.11990.

[17] 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, 2022.

[18] C. Lawrence Zitnick, Abhishek Das, Adeesh Kolluru, Janice Lan, Muhammed Shuaibi, Anuroop Sriram, Zachary Ulissi, and Brandon Wood. Spherical channels for modeling atomic interactions, 2022. URL https://arxiv.org/abs/2206.14331.

[19] Jianing Lu, Cheng Wang, and Yingkai Zhang. Predicting molecular energy using force-field optimized geometries and atomic vector representations learned from an improved deep tensor neural network. *Journal of Chemical Theory and Computation*, 15(7):4113–4121, May 2019. doi: 10.1021/acs.jctc.9b00001. URL https://doi.org/10.1021/acs.jctc.9b00001.

[20] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3D Infomax improves gnns for molecular property prediction. 2021. doi: 10.48550/ARXIV.2110.04126. URL https://arxiv.org/abs/2110.04126.

[21] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3D geometry, 2021. URL https://arxiv.org/abs/2110.07728.

[22] Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. Molecule3D: A benchmark for predicting 3D geometries from molecular graphs, 2021. URL https://arxiv.org/abs/2110.01717.

[23] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, May 2021. doi: 10.1021/acscatal.0c04525. URL https://doi.org/10.1021/acscatal.0c04525.

[24] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[25] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010. URL http://jmlr.org/papers/v11/vincent10a.html.

[26] Yingbo Zhou, Devansh Arpit, Ifeoma Nwogu, and Venu Govindaraju. Is joint training better for deep auto-encoders?, 2014. URL https://arxiv.org/abs/1405.1380.

[27] 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, 2021. URL https://arxiv.org/abs/2106.07971.

[28] 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.

[29] Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with SE(3)-invariant denoising distance matching, 2022. URL https://arxiv.org/abs/2206.13602.

[30] 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), August 2014. doi: 10.1038/sdata.2014.22. URL https://doi.org/10.1038/sdata.2014.22.

[31] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Deep learning of activation energies. *The Journal of Physical Chemistry Letters*, 11(8):2992–2997, March 2020. doi: 10.1021/acs.jpclett.0c00500. URL https://doi.org/10.1021/acs.jpclett.0c00500.

[32] Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: Using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55 (12):2562–2574, November 2015. doi: 10.1021/acs.jcim.5b00654. URL https://doi.org/10.1021/acs.jcim.5b00654.

[33] RDKit Contributors. RDKit: Open-source cheminformatics, 2021. URL http://www.rdkit.org/. Accessed: 2023-03-19.

[34] Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, April 1996. doi: 10.1002/(sici)1096-987x(199604)17:5/6<490::aid-jcc1>3.0.co;2-p. URL https://doi.org/10.1002/(sici)1096-987x(199604)17:5/6<490::aid-jcc1>3.0.co;2-p.

[35] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

[36] Kevin A. Spiekermann, Lagnajit Pattanaik, and William H. Green. Fast predictions of reaction barrier heights: Toward coupled-cluster accuracy. *The Journal of Physical Chemistry A*, 126 (25):3976–3986, June 2022. doi: 10.1021/acs.jpca.2c02614. URL https://doi.org/10.1021/acs.jpca.2c02614.

[37] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

[38] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

[39] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.

[40] Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Random features strengthen graph neural networks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 333–341. SIAM, 2021.

[41] Yaochen Xie, Zhao Xu, and Shuiwang Ji. Self-supervised representation learning via latent graph prediction. In *International Conference on Machine Learning*, pages 24460–24477. PMLR, 2022.

[42] Yaochen Xie, Zhengyang Wang, and Shuiwang Ji. Noise2Same: Optimizing a self-supervised bound for image denoising. *Advances in Neural Information Processing Systems*, 33:20320–20330, 2020.

[43] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller. SchNetPack: A deep learning toolbox for atomistic systems. *Journal of Chemical Theory and Computation*, 15(1):448–455, November 2018. doi: 10.1021/acs.jctc.8b00908. URL https://doi.org/10.1021/acs.jctc.8b00908.

[44] Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules, 2020. URL https://arxiv.org/abs/2011.14115.

[45] Limei Wang, Yi Liu, Yuchao Lin, Haoran Liu, and Shuiwang Ji. ComENet: Towards complete and efficient message passing for 3D molecular graphs. *arXiv preprint arXiv:2206.08515*, 2022.

[46] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9377–9388. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/schutt21a.html.

[47] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

[48] 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.

[49] Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve E(3) equivariant message passing, 2021. URL https://arxiv.org/abs/2110.02905.

[50] Schrödinger, LLC. The PyMOL molecular graphics system, version 2.0. 2017.

[51] Kenneth Atz, Clemens Isert, Markus N. A. Böcker, José Jiménez-Luna, and Gisbert Schneider. Δ-quantum machine-learning for medicinal chemistry. *Physical Chemistry Chemical Physics*, 24(18):10775–10783, 2022. doi: 10.1039/d2cp00834c. URL https://doi.org/10.1039/d2cp00834c.

[52] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A large-scale challenge for machine learning on graphs, 2021. URL https://arxiv.org/abs/2103.09430.

[53] Dominic Masters, Josef Dean, Kerstin Klaser, Zhiyi Li, Sam Maddrell-Mander, Adam Sanders, Hatem Helal, Deniz Beker, Ladislav Rampášek, and Dominique Beaini. GPS++: An optimised hybrid mpnn/transformer for molecular property prediction, 2022.

[54] Ferruccio Palazzesi, Markus R. Hermann, Marc A. Grundl, Alexander Pautsch, Daniel Seeliger, Christofer S. Tautermann, and Alexander Weber. BIreactive: A machine-learning model to estimate covalent warhead reactivity. *Journal of Chemical Information and Modeling*, 60(6): 2915–2923, April 2020. doi: 10.1021/acs.jcim.9b01058. URL https://doi.org/10.1021/acs.jcim.9b01058.

[55] Kevin Spiekermann, Lagnajit Pattanaik, and William H. Green. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Scientific Data*, 9(1), July 2022. doi: 10.1038/s41597-022-01529-6. URL https://doi.org/10.1038/s41597-022-01529-6.

[56] Seonghwan Kim, Jeheon Woo, and Woo Youn Kim. Diffusion-based generative AI for exploring transition states from 2D molecular graphs, 2023. URL https://arxiv.org/abs/2304.12233.

[57] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, January 1996. doi: 10.1021/jm9602928. URL https://doi.org/10.1021/jm9602928.

A Theoretical Basis

A.1 Conditionally independence

Proposition A.1. *For any random variables $(\tilde{X}, X, Y)$, if $\tilde{X}$ and Y are conditional independent given X, then $I(\tilde{X}; Y|X) = 0$.*

Proof of Proposition A.1. From the definition of conditional mutual information (MI), we start from the below.

$$
\begin{aligned}
I(\tilde{X}; Y|X) &= H(\tilde{X}|X) + H(Y|X) - H(\tilde{X}, Y|X) \\
&= \mathbb{E}_{p(\tilde{X},X,Y)}\left[-\log \frac{p(\tilde{X}|X)p(Y|X)}{p(\tilde{X}, Y|X)}\right] \\
&= \mathbb{E}_{p(\tilde{X},X,Y)}\left[-\log \frac{p(\tilde{X}, X)p(X, Y)/p(X)^2}{p(\tilde{X}, X, Y)/p(X)}\right] \\
&= \mathbb{E}_{p(\tilde{X},X,Y)}\left[-\log \frac{p(\tilde{X}, X)p(X, Y)}{p(\tilde{X}, X, Y)p(X)}\right] \\
&= \mathbb{E}_{p(\tilde{X},X,Y)}\left[-\log \frac{p(X, Y)/p(X)}{p(\tilde{X}, X, Y)/p(\tilde{X}, X)}\right] \\
&= \mathbb{E}_{p(\tilde{X},X,Y)}\left[-\log \frac{p(Y|X)}{p(Y|\tilde{X}, X)}\right] \\
&= \mathbb{E}_{p(\tilde{X},X,Y)}\left[-\log \frac{\cancel{p(Y|X)}}{\cancel{p(Y|\tilde{X}, X)}}\right] \qquad \text{(conditional independence)} \\
&= 0
\end{aligned}
$$

□

A.2 Lower bound of three-term MI

Here, we derive the lower bound of three-term MI described at Proposition 3.1.

Proof of Proposition 3.1. We need to prove the following inequality first.

$$
H(Z) - I(\tilde{Z}; Z) \geq I(Z; Y) - I(\tilde{Z}; Z; Y) \quad \forall \text{ random variables } \tilde{Z}, Z, Y
$$

$$
\begin{aligned}
\text{LHS} &= H(Z|\tilde{Z}) \\
\text{RHS} &= I(Z; Y) - [I(Z; Y) - I(Z; Y|\tilde{Z})] \\
&= I(Z; Y|\tilde{Z}) \\
&= \mathbb{E}_{p(\tilde{Z},Z,Y)}\left[-\log \frac{p(Z|\tilde{Z})p(Y|\tilde{Z})}{p(Z, Y|\tilde{Z})}\right] \\
&= \mathbb{E}_{p(\tilde{Z},Z,Y)}\left[-\log \frac{p(Z|\tilde{Z})}{p(Z|\tilde{Z}, Y)}\right] \\
&= H(Z|\tilde{Z}) - H(Z|\tilde{Z}, Y)
\end{aligned}
$$

Since conditional entropy is non-negative, LHS≥RHS.

By applying the above, we derive the following two inequalities:

$$
\begin{aligned}
I(\tilde{Z}; Z; Y) &\geq I(\tilde{Z}; Z) + I(Z; Y) - H(Z) \\
&= -[H(Z) - I(\tilde{Z}; Z)] - [H(Y) - I(Z; Y)] + H(Y) \\
&= -H(Z|\tilde{Z}) - H(Y|Z) + H(Y),
\end{aligned}
$$

$$
\begin{aligned}
I(\tilde{Z}; Z; Y) &\geq I(\tilde{Z}; Y) + I(Z; Y) - H(Y) \\
&= -[H(Y) - I(\tilde{Z}; Y)] - [H(Y) - I(Z; Y)] + H(Y) \\
&= -H(Y|\tilde{Z}) - H(Y|Z) + H(Y).
\end{aligned}
$$

By adding the two inequalities, we derive a lower bound:

$$
I(\tilde{Z}; Z; Y) \geq H(Y) \underbrace{-H(Y|Z) - \frac{1}{2}H(Y|\tilde{Z}) - \frac{1}{2}H(Z|\tilde{Z})}_{\text{LB}}.
$$

□

Though the coefficient of the lower bound is different for each term, our practical loss is calculated as follows:

$$
\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{y,corrupted}} + \mathcal{L}_{\text{y,correct}} + \lambda \mathcal{L}_d,
$$

where $\mathcal{L}_{\text{y,corrupted}}$, $\mathcal{L}_{\text{y,correct}}$, and $\mathcal{L}_d$ correspond to $H(Y|\tilde{Z})$, $H(Y|Z)$, and $H(Z|\tilde{Z})$, respectively, and the coefficient λ is adopted for a practical reason. The searching space of λ is described in Appendix C.

A.3 Choice of denoising loss

Surrogate loss. We decode X from $\tilde{Z}$ as a surrogate task for $H(Z|\tilde{Z})$. Since an MI is invariant to continuous and bijective mappings, the surrogate loss to reconstruct X can be obtained by transforming from continuous representation space to data space. Although general 3D GNNs do not satisfy this requirement, we have empirically confirmed robust results in various chemistry tasks using several GNNs.

Nevertheless, it is necessary to explain the difference between maximizing $I(\tilde{Z}; Z)$ and $I(X|\tilde{Z})$. As shown in Figure 1(A), various properties can be obtained from X, implying that X contains information that is necessary to predict all the properties. However, Z partially contains the information of X in that it is a representation of X. For a specific property prediction task, the ideal situation would be for Z to contain only the information necessary to accurately predict Y. In this context, mapping $\tilde{Z}$ to X instead of Z may cause superfluous information which is irrelevant to Y. But even so, it does no harm for our overall purpose of addressing the inductive bias of physical relationship.

Geometric denoising loss. We incorporated a geometric denoising loss as a surrogate loss in order to maximize the MI. Specifically, we aimed to maintain the equivariance of X under rotation or translation of $\tilde{X}$. To achieve this, we employed the SE(3)-equivariant decoders in this study.

For the loss metric, we chose the MAE of the atom-pair distances. This choice was natural considering the importance of bond distances in molecular geometry compared to absolute atomic positions. Specifically, the denoising metric $\mathcal{L}$ is calculated based on the (i, j) atom-pair distances d_{ij} and $\tilde{d}_{ij}$ of x and $\tilde{x}$, respectively, which is

$$
\mathcal{L}\left(x, \tilde{x}\right) = \frac{1}{|\mathcal{E}|} \sum_{(i,j) \in \mathcal{E}} |d_{ij} - \tilde{d}_{ij}|. \tag{2}
$$

Here, $\mathcal{E}$ denotes a set of edges in a graph of $\tilde{x}$. Thus, the practical denoising loss is calculated as follows:

$$
\mathcal{L}_d = \frac{1}{|\mathcal{D}|} \sum_{(\tilde{x}, x, y) \in \mathcal{D}} \mathcal{L}\big(x, g\left(f\left(\tilde{x}\right)\right)\big), \tag{3}
$$

where f and g denote an encoder and a decoder, respectively, and $\mathcal{D}$ is the dataset. However, in case of the OC20 dataset, we computed the loss in atomic positions to align with the baseline model for comparisons.

To effectively minimize the denoising loss, we introduced an explicit position update at each layer, inspired by the geometry optimization of quantum chemical calculation. These position update layers facilitated the differentiation between the encoders of $\tilde{X}$ and X, despite sharing model parameters. It involved incorporating relatively small parameter additions to induce distinct mappings.

To achieve effective denoising and align the encoding process with the geometry optimization, we introduced a gradual denoising loss. This additional loss term guides the position updates at each layer to exhibit directional behavior by forcing the updated geometry to lie within the linear interpolation between $\tilde{X}$ and X. We compared the prediction performance according to the degree of corruption of the input geometry, and confirmed that the prediction performance improved as the geometry closer to X was used (see Appendix B.1). It indirectly explains the reason for the introducing explicit position update and gradual loss. More details on the loss form and ablation studies of gradual denoising are provided in Appendix B.2.

B Further analyses

B.1 Utilization of interpolated geometries

In this study, we have assumed a Markov chain given that the correct geometry X is optimized from the corrupted geometry $\tilde{X}$, and the quantum chemical property Y is computed from X. Since X is more adjacent to Y in the Markov chain, predictions of Y based on X should be more accurate than those based on $\tilde{X}$. Furthermore, from a physical standpoint, the transition from $\tilde{X}$ to X is a form of geometry optimization that could be perceived as a Markov chain. Building on these assumptions, the intermediate geometry during the optimization process naturally lies between X and $\tilde{X}$ within the Markov chain.

In this section, we explored the possibility of using interpolated geometry $\frac{X+\tilde{X}}{2}$ as a surrogate intermediate geometry. We expected that the following two statements will be satisfied.

- Regardless of the types of properties, the baseline (trained with $\tilde{X}$) will always have a higher MAE than the ground truth (trained with X).
- The interpolated geometry between $\tilde{X}$ and X is a less corrupted geometry than $\tilde{X}$. Thus, MAEs of the model using the interpolated geometry are placed between those of the baseline and the ground truth.

For the demonstration, we used three individual prediction models, where the training input for each model was X, $\frac{X+\tilde{X}}{2}$, and $\tilde{X}$, respectively. As expected, the results presented in Tables 5 and 6 consistently show that the predictive accuracy of the model decreases as the level of corruption in the geometry increases. This empirical evidence underscores the superiority of interpolated geometry over $\tilde{X}$ as an input for predicting various quantum chemical properties.

Thus, we can consider $\tilde{X} \rightarrow \frac{X+\tilde{X}}{2} \rightarrow X$ as a proxy of the geometry optimization process which is a Markov chain. Inspired by these findings, we integrated the explicit position update scheme of EGNN [15] and introduced an additional loss that exploits the interpolated geometry in the position update step following geometry optimization.

Table 5: MAEs for QM9's properties according to different inputs. Values were obtained using EGNN. ($\tilde{X}$+X)/2 denotes the interpolated geometry generated from the mean of the atom-pair distance of $\tilde{X}$ and X.

Target	Unit	$\tilde{X}$	($\tilde{X}$+X)/2	X
U_0	meV	17.4	14.2	12.9
μ	D	0.133	0.0807	0.0350
α	Bohr^3	0.125	0.0947	0.0759
ϵ_{HOMO}	meV	38.4	33.4	31.2
ϵ_{LUMO}	meV	34.4	28.9	26.6
GAP	meV	58.0	53.0	51.1
R^2	Bohr^2	5.60	2.92	0.130
C_v	$\text{cal/mol} \cdot \text{K}$	0.0445	0.0377	0.0336
ZPVE	meV	1.97	1.74	1.59

Table 6: MAEs (kcal/mol) of predicted barrier heights according to different inputs. Values were obtained using DimeReaction. ($\tilde{X}$+X)/2 denotes the pairwise interpolated geometry generated from the mean of the atom-pair distance of (X^R, X^P) and (X^R, X^{TS}).

Dataset	$\tilde{X}$	($\tilde{X}$+X)/2	X
CCSD(T)-UNI	6.49	5.30	2.38
B97-D3	8.24	3.91	1.92

B.2 Ablation study for denoising process

In GeoTMI, the denoising process proceeds throughout all GNN layers. The relationship between the number of GNN layers and the average absolute difference based on the atom-pair distance (D-MAE) between the correct geometry and the predicted geometry of the atomic positions at the last layer was investigated using the EGNN model (see Figure 2). We note that each EGNN was trained using only the denoising loss to confirm the denoising power alone.

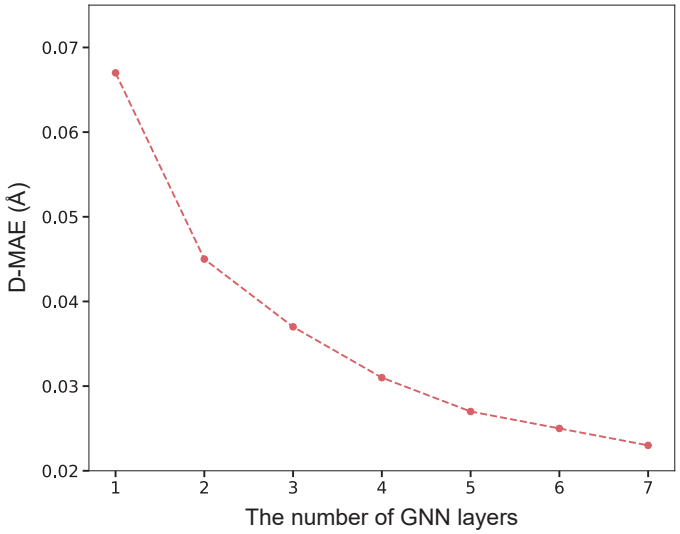

Figure 2: D-MAE according to the number of GNN layers of the EGNN. The D-MAE is measured using X and the denoised $\tilde{X}$ derived from the last layer.

In particular, the denoising process with the small number of GNN layers could not restore X. We note that the D-MAE between the MMFF and the DFT geometries is 0.0712Å. If the denoising process couldn't restore X, the corresponding denoising loss can have a negative effect on learning by maintaining a large amount compared to other losses in the training process. In this respect, we designed a gradual denoising loss, which is a slightly modified version of Equation (3) with a linearly interpolated target instead of x. For each position update layer, a target varies linearly from $\tilde{x}$ to x, and the loss is calculated according to Equation (2). Specifically, the target distance of the l-th layer among a total of L layers is as follows:

$$\bar{d}_{ij}^{l} = \frac{1}{L}\left(ld_{ij} + (L-l)\tilde{d}_{ij}\right).$$

Table 7 shows that the gradual denoising process is useful to increase model performance. Thus, we adopted the gradual denoising loss to predict quantum chemical properties.

B.3 Application of GeoTMI on SchNet and DimeNet++ for QM9 and QM9$_{\mathrm{M}}$ tasks

GeoTMI is applicable regardless of the 3D geometry model architectures. In this regard, we tested the effectiveness of GeoTMI using two additional 3D GNNs: SchNet [43] and DimeNet++ [44]. For

Table 7: Impact of each denoising task in terms of MAE. We tested gradual denoising ("GeoTMI"), "w/o Gradual denoising", and "Denoising last only" for four properties. When the gradual denoising was not used, the objective of all denoising processes of GNN layers is restoring X. The "Denoising last only" means that denoising objective contains only the denoising loss of the last GNN layer without the losses of other GNN layers.

Property	Unit	GeoTMI	w/o Gradual denoising	Denoising last only
U_0	meV	14.5	15.4	14.9
R^2	Bohr^2	4.08	4.22	4.97
C_v	$\mathrm{cal/mol \cdot K}$	0.0407	0.0413	0.0423
μ	D	0.0997	0.100	0.132

comparison, we also identified the performance of QM9's properties prediction from X for the two models (see Table 8). The training, validation, and testing data were used as in Section 4.1.

Table 8: MAEs for QM9's properties. GeoTMI was tested with two different 3D GNNs: SchNet and DimeNet++.

Approach	Input type (Train / Infer.)	U_0 (meV)	μ (D)	α (Bohr^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	GAP (meV)	R^2 (Bohr^2)	C_v ($\frac{cal}{mol \cdot K}$)	ZPVE (meV)
SchNet	X/X	17.0	0.0391	0.0859	44.3	34.9	68.6	0.170	0.0313	1.67
DimeNet++	X/X	8.99	0.0382	0.0583	25.5	20.8	43.0	0.342	0.0255	1.30
SchNet	$\tilde{X}/\tilde{X}$	27.3	0.208	0.160	61.4	53.9	85.8	8.32	0.0595	2.57
SchNet + GeoTMI	$X, \tilde{X}/\tilde{X}$	27.3	0.139	0.131	52.0	45.0	74.4	6.44	0.0566	2.38
DimeNet++	$\tilde{X}/\tilde{X}$	16.9	0.140	0.123	37.5	35.1	56.3	5.82	0.0462	2.10
DimeNet++ + GeoTMI	$X, \tilde{X}/\tilde{X}$	14.0	0.127	0.109	35.0	32.6	55.3	5.29	0.0418	1.90
Improvements in SchNet (%)		0.00	33.2	18.1	15.3	16.5	13.3	22.6	4.87	7.39
Improvements in DimeNet++ (%)		17.2	9.29	11.4	6.72	7.12	1.78	9.11	9.52	9.52

B.4 Performance of GeoTMI based on OOD data for QM9's properties

We identified the out-of-distribution (OOD) generalization ability of GeoTMI using the EGNN model on QM9's properties. To this end, we arranged 100,000, 18,000, and 13,000 molecules for training, validation, and testing, respectively, based on a scaffold split, ensuring that the molecules in the OOD were included in the test set. The split is based on the Bemis–Murcko scaffold [57] implemented in RDKit [33] library. Table 9 shows that GeoTMI has improved the model prediction performance regardless of OOD data for the tested properties. The results show the robustness of GeoTMI in terms of OOD generalization ability.

Table 9: The MAEs for R^2, μ, and U_0 in the $\mathrm{QM9_M}$. We verified the performance of GeoTMI on testing datasets using random and scaffold splits, respectively. For the MAE of each property, the same units are used as in Table 1.

Split	Approach	μ	R^2	U_0
Random	EGNN	0.133	5.60	17.4
	EGNN + GeoTMI	0.100	4.08	14.5
Scaffold	EGNN	0.195	10.4	33.0
	EGNN + GeoTMI	0.149	7.60	23.4

C Experimental details

C.1 Parameter details

QM9$_{\mathrm{M}}$. We used the reported hyperparameters optimized for QM9 from previous studies for EGNN [15], SchNet [43], DimeNet++ [44], and Transformer-M [17], respectively. The search space of λ is specified in Table 10.

Table 10: The search space of λ on $\mathrm{QM9}_{\mathrm{M}}$ task.

Target	EGNN	DimeNet++	SchNet
U_0	[0.1, 0.5, 1.0, 10.0]	0.1	[0.1, 0.5, 1.0, 5.0, 10.0]
μ	[0.1, 0.5, 1.0, 10.0]	0.1	[0.1, 0.5, 1.0, 5.0, 10.0]
α	0.1	0.1	[0.1, 0.5, 1.0, 5.0, 10.0]
ϵ_{HOMO}	0.1	0.1	0.1
ϵ_{LUMO}	0.1	0.1	0.1
GAP	0.1	0.1	0.1
R^2	[0.1, 0.5, 1.0, 10.0]	0.1	[0.1, 0.5, 1.0, 5.0, 10.0]
C_v	0.1	0.1	0.1
ZPVE	0.1	0.1	0.1

Reaction barrier prediction. For both DimeReaction models with and without GeoTMI, we used the same hyperparameters as Spiekermann et al. [55] except for the number of epochs and batch sizes. Table 11 shows the values of the number of epochs and batch sizes used in this work. The search space of λ is specified in Table 12.

Table 11: The hyperparameters used for training DimeReaction.

Parameter	CCSD(T)-UNI	B97-D3
Epoch	200	200
Batch size	32	64
Warm-up epochs	3	3

Table 12: The search space of λ for CCSD(T)-UNI and B97-D3 datasets.

Parameter	CCSD(T)-UNI	B97-D3
λ	[0.005, 0.01, 0.05, 0.1, 0.5, 1.0]	[0.005, 0.01, 0.05, 0.1, 0.5, 1.0]