Pre-training of Molecular GNNs as Conditional Boltzmann Generator

Anonymous Submission

Abstract

Learning representations of molecular structures using deep learning is a fundamental problem in molecular property prediction tasks. Molecules inherently exist in the real world as three-dimensional structures; furthermore, they are not static but in continuous motion in the 3D Euclidean space, forming a potential energy surface. Therefore, it is desirable to generate multiple conformations in advance and extract molecular representations using a 4D-QSAR model that incorporates multiple conformations. However, this approach is impractical for drug and material discovery tasks because of the computational cost of obtaining multiple conformations. To address this issue, we propose a pre-training method for molecular GNNs using an existing dataset of molecular conformations to infer a latent vector universal to multiple conformations from a 2D molecular graph. Our method, called Boltzmann GNN, is formulated by maximizing the conditional marginal likelihood of a conditional generative model for conformations generation. We show that our model has a better prediction performance for molecular properties than existing pre-training methods using molecular graphs and three-dimensional molecular structures.

1. Introduction

Learning representations of molecular structures using deep learning is a useful approach in drug and material discovery (G**** et al. 2018; S***** et al. 2020; Z*** et al. 2018). In particular, for the task of molecular property prediction, Graph Neural Networks (GNNs) have been successful (G***** et al. 2018; D***** et al. 2015; F**** et al. 2020). Various models of these architectures have been studied, depending on the prediction task. For example, a Graph Field Network (GFN) can predict the potential energy of a molecule from the coordinates of its atomic nucleus (S***** et al. 2017; S**** et al. 2018). Although GNNs and GFNs treat molecules as stationary objects, to accurately predict biological or physico-chemical properties, we should use their conformation ensembles. This is because molecules are not static but are in continuous motion in 3D Euclidean space, forming a potential energy surface (PES) (S**** et al. 2003; H**** et al. 2017). Molecular chemical properties are a function of the set of conformations (conformation ensemble) accessible at a finite temperature (K*** et al. 2016).

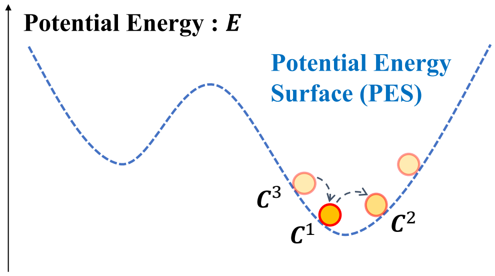

Figure 1 : Illustration of conformations on potential energy surface. The coordinate geometry $\boldsymbol{C^i}$ on the PES is called conformation.

Figure 1 shows the relationship between the PES and conformations.

Recent studies (Z**** et al. 2021; W**** et al. 2021; A**** et al. 2023) used molecular dynamics (MD) simulations to generate conformation ensembles and used the conformation ensemble as an input to a DNN to predict molecular properties. Here, we refer to these models as 4D-QSARs. These approaches make sense from a physical perspective. However, the use of classical molecular dynamics simulations to explicitly compute a conformation ensemble before predicting its properties is computationally intractable for many real-world applications.

In this study, we propose a pre-training method for GNNs using an existing dataset of conformation ensembles as a surrogate model for 4D-QSARs. From the perspective of statistical physics, a conformation $\boldsymbol{C}$ can be treated as a random quantity sampled from the Boltzmann distribution $p^*(\boldsymbol{C}) \propto \exp(-E(\boldsymbol{C}))$, where $E(\boldsymbol{C})$ is the potential energy of $\boldsymbol{C}$. If we can obtain a conformation ensemble on the PES as observed samples that follow the $p^*(\boldsymbol{C})$, we can estimate a universal latent vector for multiple conformations using a conditional generative model.

1.1 Related Works

Pre-training methods for Molecular GNNs using molecular conformations have been proposed to obtain a better prediction performance for molecular properties. GraphMVP

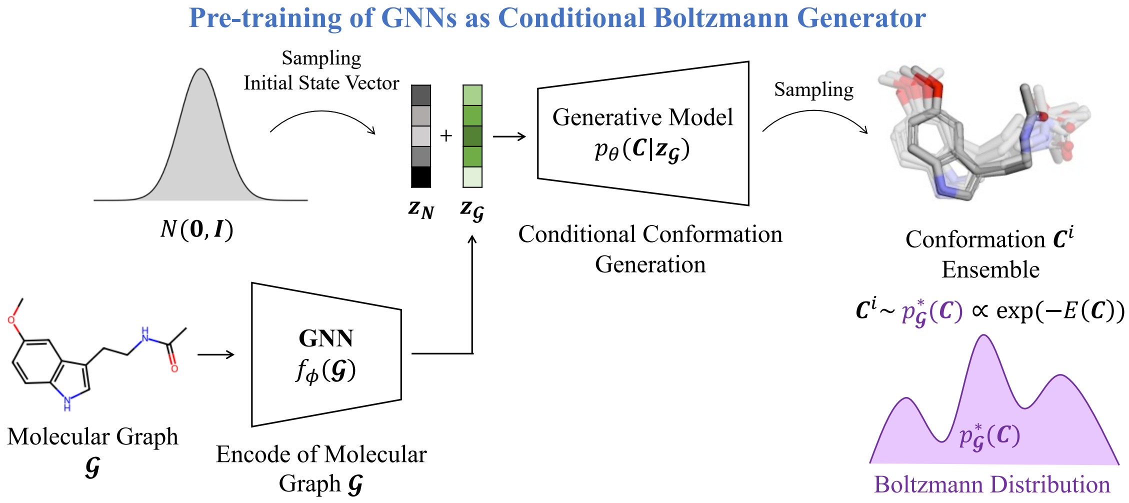
Pre-training of GNNs as Conditional Boltzmann Generator
Sampling
Initial State Vector
$N(\mathbf{0}, \boldsymbol{I})$
z_N
$z_{\mathcal{G}}$
Generative Model
$p_\theta(\boldsymbol{C}|\boldsymbol{z}_{\mathcal{G}})$
Conditional Conformation Generation
Sampling
Conformation $\boldsymbol{C}^i$ Ensemble
$\boldsymbol{C}^i \sim p_{\mathcal{G}}^*(\boldsymbol{C}) \propto \exp(-E(\boldsymbol{C}))$
$p_{\mathcal{G}}^*(\boldsymbol{C})$
Boltzmann Distribution
GNN
$f_\phi(\mathcal{G})$
Molecular Graph $\mathcal{G}$
Encode of Molecular Graph $\mathcal{G}$

Figure 2 : Scheme of our pre-training method for molecular GNNs. Our pre-training method generates a conditional variable $\boldsymbol{z}_{\mathcal{G}}$ for the conditional generative model $\boldsymbol{p}_{\boldsymbol{\theta}}(\boldsymbol{C}|\boldsymbol{z}_{\mathcal{G}})$, which generates a molecular conformation ensemble using a molecular GNN that encodes a molecular graph $\mathcal{G}$ into the latent vector $\boldsymbol{z}_{\mathcal{G}}$. $\boldsymbol{p}_{\boldsymbol{\theta}}(\boldsymbol{C}|\boldsymbol{z}_{\mathcal{G}})$ approximates $\boldsymbol{p}_{\mathcal{G}}^*(\boldsymbol{C})$.

(L**** et al. 2021) proposes a generative and a contrastive learning task for maximizing between molecular 2D topologies and the 3D conformations. 3D Infomax (S**** et al. 2022) proposes a knowledge distillation method from a GFN using molecular 3D geometries to GNNs using molecular graphs. These methods use information from multiple conformations but do not explicitly incorporate information on the Boltzmann distribution.

We introduce a pre-training method for molecular GNNs to incorporate information on the Boltzmann distribution.

2. Methods

2.1 Preliminaries

3D conformation of molecule. For geometry, each atom a_i in molecule $\mathcal{M}$ is embedded by a coordinate vector $\boldsymbol{c}_i \in \mathbb{R}^3$ into 3D space, and the full set of positions (conformation) can be represented as a matrix $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \dots \boldsymbol{c}_n] \in \mathbb{R}^{n\times 3}$.

2D molecular graph. A molecular graph is denoted as $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{\boldsymbol{a}_i\}_{i=1}^{n}$ is the set of vertices representing atoms and $\boldsymbol{E} = \{\boldsymbol{b}_{ij} \mid (i,j) \subseteq |\boldsymbol{V}| \times |\boldsymbol{V}|\}$ is the set of edges representing the inter-atomic bonds.

2.2 Motivation

Let $p_{\mathcal{G}}^*(\boldsymbol{C})$ be the Boltzmann distribution of the conformation ensemble $(\boldsymbol{C}^1, \boldsymbol{C}^2, \dots \boldsymbol{C}^N)$ for a molecular graph $\mathcal{G}$. Our aim is to obtain the latent vector $\boldsymbol{z}_{\mathcal{G}}$ of the molecular graph $\mathcal{G}$ to be used as the conditional variable for a conditional generative model $p_\theta(\boldsymbol{C}|\boldsymbol{z}_{\mathcal{G}})$ to approximate the $p_{\mathcal{G}}^*(\boldsymbol{C})$. This problem is to obtain $\boldsymbol{z}_{\mathcal{G}}^* \in \mathbb{R}^D$ such that,

$$\boldsymbol{z}_{\mathcal{G}}^* = \underset{\boldsymbol{z}_{\mathcal{G}} \in \mathbb{R}^D}{\operatorname{argmin}}\, KL\left[p_{\mathcal{G}}^*(\boldsymbol{C})||p_\theta(\boldsymbol{C}|\boldsymbol{z}_{\mathcal{G}})\right], \qquad (1)$$

where θ is a set of parameters for conditional generative model $p_\theta(\boldsymbol{C}|\boldsymbol{z}_{\mathcal{G}})$. If N is large, $KL[\cdot]$ of Eq. 1 can be rewritten as follows:

$$\widehat{\boldsymbol{z}_{\mathcal{G}}} = \underset{\boldsymbol{z}_{\mathcal{G}} \in \mathbb{R}^D}{\operatorname{argmin}} \left[\frac{1}{N}\left(-\sum_{i=1}^{N} \log\left(p_\theta(\boldsymbol{C}^{\boldsymbol{i}}|\boldsymbol{z}_{\mathcal{G}})\right)\right) + H\left(p_{\mathcal{G}}^*(\boldsymbol{C})\right)\right]. \quad (2)$$

$H\left(p_{\mathcal{G}}^*(\boldsymbol{C})\right)$ is the entropy of $p_{\mathcal{G}}^*(\boldsymbol{C})$ and constant term. Therefore, Eq. 3 represents the maximum likelihood estimator (MLE). This maximum likelihood estimate $\widehat{\boldsymbol{z}_{\mathcal{G}}}$ includes information on the conformation ensemble $(\boldsymbol{C}^1, \boldsymbol{C}^2, \dots \boldsymbol{C}^N)$ on the PES. In practice, the latent vector encoded by a $f_\phi(\mathcal{G})$ (GNN) is denoted as $\boldsymbol{z}_{\mathcal{G}}$, and we estimate the parameter $\hat{\phi}$ of the $f_\phi(\mathcal{G})$ using MLE. Therefore, we can obtain the following objective：

$$\hat{\phi} = \underset{\phi \in \mathbb{R}^d}{\operatorname{argmin}} \left[\frac{1}{N}\left(-\sum_{i=1}^{N} \log\left(p_\theta(\boldsymbol{C}^{\boldsymbol{i}}|f_\phi(\mathcal{G}))\right)\right)\right]. \qquad (3)$$

We use this MLE as an objective function for training GNNs. Our goal is to improve the prediction performance of molecular properties for small datasets using this training method for the pre-training of GNNs.

Figure 2 shows a schematic illustration of our solution, which constructs a conditional Boltzmann generator that approximates the Boltzmann distribution using a hierarchical model of a GNN and conditional generative model.

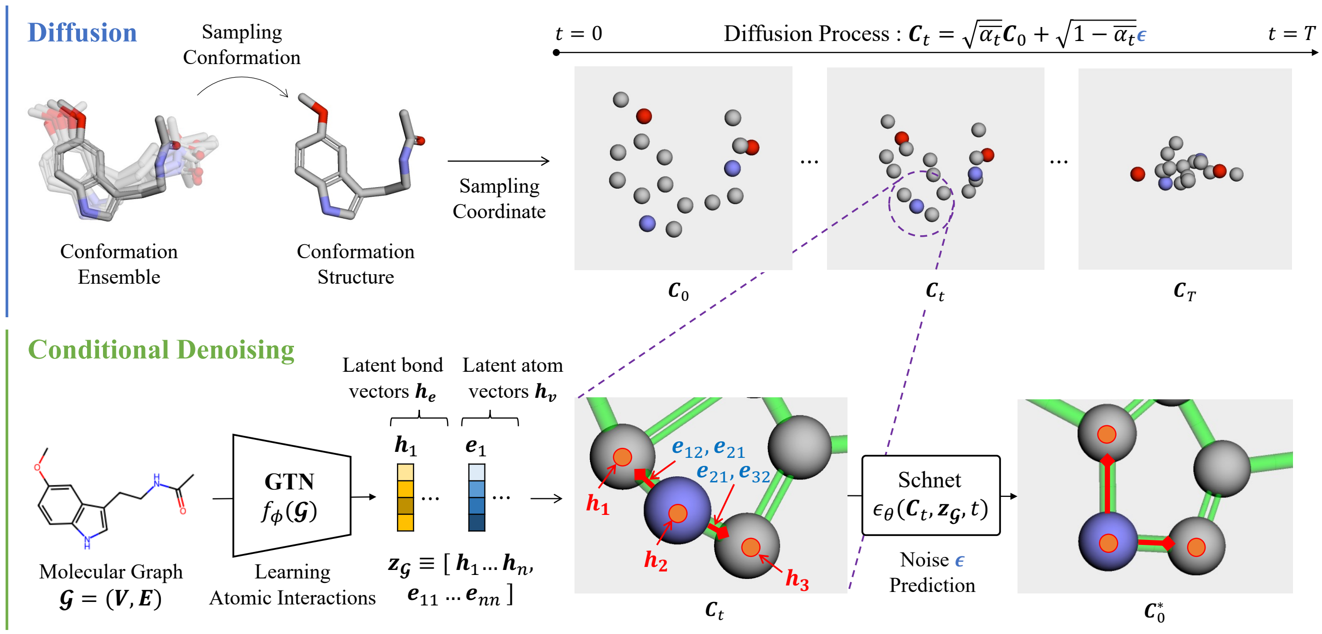
Diffusion
Sampling Conformation
$t = 0$
Diffusion Process : $\boldsymbol{C}_t = \sqrt{\overline{\alpha_t}}\boldsymbol{C}_0 + \sqrt{1-\overline{\alpha_t}}\boldsymbol{\epsilon}$
$t = T$
Conformation Ensemble
Conformation Structure
Sampling Coordinate
...
...
$\boldsymbol{C}_0$
$\boldsymbol{C}_t$
$\boldsymbol{C}_T$
Conditional Denoising
Latent bond vectors $\boldsymbol{h}_e$
Latent atom vectors $\boldsymbol{h}_v$
$\boldsymbol{h}_1$
$\boldsymbol{e}_1$
GTN
$f_\phi(\mathcal{G})$
Molecular Graph
$\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$
Learning Atomic Interactions
$\boldsymbol{z}_{\mathcal{G}} \equiv [\,\boldsymbol{h}_1 \dots \boldsymbol{h}_n,$
$\boldsymbol{e}_{11} \dots \boldsymbol{e}_{nn}\,]$
$\boldsymbol{e}_{12}, \boldsymbol{e}_{21}$
$\boldsymbol{e}_{21}, \boldsymbol{e}_{32}$
$\boldsymbol{h}_1$
$\boldsymbol{h}_2$
$\boldsymbol{h}_3$
$\boldsymbol{C}_t$
Schnet
$\epsilon_\theta(\boldsymbol{C}_t, \boldsymbol{z}_{\mathcal{G}}, t)$
Noise $\boldsymbol{\epsilon}$ Prediction
$\boldsymbol{C}_0^*$

Figure 3 : Illustration of our pre-training method. First, the diffusion process adds a noise $\boldsymbol{\epsilon}$ to the coordinates of a conformation $\boldsymbol{C}_{\mathbf{0}}$. GTN encodes a molecular graph $\mathcal{G}$ to the latent vector $\boldsymbol{z}_{\mathcal{G}}$ for embedding atomic interactions, and Schnet uses it to predict the noise $\boldsymbol{\epsilon}$. $\boldsymbol{C}_{\mathbf{0}}^*$ indicates the original conformation before noise is added.

2.3 Model Architecture

Conditional generative model. Conditional generative model $p_\theta(\boldsymbol{C}|\boldsymbol{z}_{\mathcal{G}})$ needs to satisfy SE(3)-invariant likelihood. Furthermore, it should be possible to model multimodal distributions, such as the Boltzmann distribution. To satisfy these requirements, we use a geometric diffusion model (Geodiff) (X**** et al. 2022).

Geodiff is a conditional generative model using denoising diffusion probabilistic model (DDPM) (H**** et al. 2020). The training process of Geodiff is to predict Gaussian noise ε from a molecular graph $\mathcal{G}$ and a noisy conformation $\boldsymbol{C}_t$ that contains a Gaussian noise $\boldsymbol{\varepsilon}_t \in \mathbb{R}^{n\times 3} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I})$. This noisy conformation is obtained by a discrete Markov chain, called the diffusion process, using the following equation:

$$\boldsymbol{C}_t = \sqrt{1-\beta_t}\boldsymbol{C}_{t-1} + \sqrt{\beta_t}\boldsymbol{\varepsilon}_t \,, \tag{4}$$

where t represents time, and $\boldsymbol{C}_0 = \boldsymbol{C}$. By increasing β_t from 0 to 1 as t increases, $\boldsymbol{C}_0$ is converted to a random noise vector. Let $\alpha_t \equiv 1 - \beta_t$ and $\overline{\alpha_t} \equiv \prod_{s=1}^{t} \alpha_s$, we get a sample $\boldsymbol{C}_t$ with noise $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 3} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I})$ from following equation:

$$\boldsymbol{C}_t = \sqrt{\overline{\alpha_t}}\boldsymbol{C}_0 + \sqrt{1-\overline{\alpha_t}}\boldsymbol{\epsilon} \,. \tag{5}$$

The objective of Geodiff is to minimize the following equation as a variational upper bound on the negative log marginal likelihood $-\log p_\theta(\boldsymbol{C}|\mathcal{G})$:

$$\mathbb{E}_{t\sim\text{Uniform}(\{1,T\})}[\|\boldsymbol{\epsilon} - \epsilon_\theta(\boldsymbol{C}_t, \mathcal{G}, t)\|^2] \,. \tag{6}$$

ϵ_θ is Schnet (S**** et al. 2018). In Geodiff, although the input $\mathcal{G}$ of Schnet is fixed, we change it to a latent vector $\boldsymbol{z}_{\mathcal{G}}$ with $\mathcal{G}$ encoded by GNNs ($f_\phi : \mathcal{G} \to \boldsymbol{z}_{\mathcal{G}}$). Thus, we change Eq. 6 to follows:

$$\mathbb{E}_{t\sim\text{Uniform}(\{1,T\})}\left[\left\|\boldsymbol{\epsilon} - \epsilon_\theta\left(\boldsymbol{C}_t, \boldsymbol{z}_{\mathcal{G}}, t\right)\right\|^2\right]. \tag{7}$$

GNN model for encoding molecular graphs. We use Graph transformer network (GTN) (D**** et al. 2020) for encoding molecular graphs into their latent vectors $\boldsymbol{z}_{\mathcal{G}}$. GTN uses self-attention (V**** et al. 2017) and Laplacian encoding to embed atomic interactions in a molecular graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ into latent atomic vectors $\boldsymbol{h}_v = \{\boldsymbol{h}_i\}_{i=1}^{n}$ and latent edge vectors $\boldsymbol{h}_e = \{\boldsymbol{e}_{ij} \mid (i,j) \subseteq |\boldsymbol{V}| \times |\boldsymbol{V}|\}$. We define $\boldsymbol{z}_{\mathcal{G}}$ as $[\boldsymbol{h}_v, \boldsymbol{h}_e]$. The encoding of GTN ($f_\phi : \mathcal{G} \to \boldsymbol{z}_{\mathcal{G}}$) is

$$\boldsymbol{h}_v, \boldsymbol{h}_e = f_\phi(\mathcal{G}) \,. \tag{8}$$

Figure 4 shows our method using GTN and Geodiff. Schnet ϵ_θ, which predicts a noise $\boldsymbol{\epsilon}$ added to the conformation $\boldsymbol{C}$, is called score function. The function estimates the molecular force field (Z**** et al. 2022). ϵ_θ estimates molecular force fields using latent vectors $\boldsymbol{z}_{\mathcal{G}}$ from $f_\phi(\mathcal{G})$, therefore, $f_\phi(\mathcal{G})$ learns latent vectors $\boldsymbol{z}_{\mathcal{G}}$ about atomic interactions on the conformation. Latent edge vectors $\boldsymbol{h}_e$ are important for extracting the atomic interactions between atoms.

Table 1: Results for molecular property prediction tasks. For each downstream task, we report the mean squared error (MSE) of 3 seeds with scaffold splitting. The best performance for each task is marked in **bold**. BACE1 and CTSD are binding affinity datasets for a target protein from Excape-DB. We wrote the sample size next to the dataset name.

Pre-training method	Small Datasets (Sample size)				
	Solubility (1.1k)	Malaria (10k)	Lipophilicity (4.2k)	BACE1 (3.6k)	CTSD (1.1k)
GraphCL	1.2189	1.2152	0.5708	0.8409	0.8066
AttrMask	1.3396	1.2522	0.5437	0.7672	0.7268
3D Infomax	1.2276	1.2263	0.5389	0.7882	0.9327
GraphMVP	1.1719	1.1619	**0.5088**	0.8058	0.7689
Boltzmann GNN	**0.8649**	**1.1586**	0.6346	**0.5984**	**0.7257**

2.4 Loss Function

The loss function is the mean of the evidence upper bound on the $-\log p_\theta(\boldsymbol{C}^{\boldsymbol{i}} | f_\phi(\boldsymbol{\mathcal{G}}))$:

$$\frac{1}{N}\left(\sum_{i=1}^{N}\mathbb{E}_{t\sim\mathrm{Uniform}(\{1,T\})}\left[\left\|\boldsymbol{\epsilon}-\epsilon_\theta\left(\boldsymbol{C}_t^i, f_\phi(\boldsymbol{\mathcal{G}}), t\right)\right\|^2\right]\right), \qquad (9)$$

where $\boldsymbol{C}_t^i$ is the conformation corresponding to the molecular graph $\boldsymbol{\mathcal{G}}$, and is the noisy conformation obtained from time t of the diffusion process. In practice, we compute the above loss Eq. 9 for the various molecular graphs and conformation ensembles in a dataset. We minimize the following expectation $\mathcal{L}(\theta,\phi)$ as in Geodiff:

$$\mathbb{E}_{(\boldsymbol{C}^i,\boldsymbol{\mathcal{G}})\sim\pi(\boldsymbol{C},\boldsymbol{\mathcal{G}}),t\sim\mathrm{Uniform}(\{1,T\})}\left[\left\|\boldsymbol{\epsilon}-\epsilon_\theta\left(\boldsymbol{C}_t^i, f_\phi(\boldsymbol{\mathcal{G}}), t\right)\right\|^2\right].$$

$\pi(\boldsymbol{C},\boldsymbol{\mathcal{G}})$ is a joint distribution of molecular graphs and conformations obtained from a dataset. We can optimize θ and ϕ with stochastic gradient descent for $\mathcal{L}(\theta,\phi)$. We call GNNs using this objective Boltzmann GNN.

3. Experiment and Results

We empirically evaluate our model with transfer learning tasks for small datasets. We compare the performance of the Boltzmann GNN with existing pre-training methods.

3.1 Setup

Datasets. For the pre-training datasets, we take 60k molecules from GEOM (A**** et al. 2022). We took 5 conformers for each molecule. For downstream tasks, we obtained datasets of binding affinity to target proteins obtained from Excape-DB (S**** et al. 2017) and datasets of physicochemical properties such as solubility and lipophilicity from MoleculeNet (W**** et al. 2018). Finally, we set five regression tasks.

Baselines. For 2D graph-based pre-training methods, we chose well-acknowledged SSL methods: GraphCL (Y** et al. 2020) and AttrMask (H** et al. 2019). For 3D structure informed pre-training methods, we chose recent proposed SSL methods: GraphMVP (L** et al. 2021) and 3D Infomax (S**** et al. 2022). We used Graph Isomorphism Network (GIN) (X** et al. 2018) and Schnet as baseline models.

Pre training.
We trained our model and baseline models for 500 epochs and determined the best model for each on validation samples not used for training.

3.2 Results

We summarized in Table 1 the mean squared error for each model in each dataset. Boltzmann GNN achieved state-of-the-art performance for four of the five tasks. Furthermore, our model performed better for datasets with small sample sizes such as Solubility and BACE1.

4. Conclusion and Future work

In this study, we proposed a novel pre-training method for molecular GNNs via conditional Boltzmann generator. We integrated the geometric diffusion model and the graph transformer to infer the latent vector of the Boltzmann distribution. Our pre-training method explicitly incorporated information on Boltzmann distribution, which improved prediction performance for downstream tasks such as molecular properties.

These results support the effectiveness of the pre-training method with conditional Boltzmann generation, and we will continue to explore further in this direction.

5. Acknowledgement

This work was supported by JSPS KAKENHI (Grant number 22J11040).

References

A *** (anonymous), 2022. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185. doi.org/10.1038/s41597-022-01288-4.

A *** (anonymous), 2023. Molecular machine learning with conformer ensembles. *Machine Learning: Science and Technology*, 4(3): 035025. doi.org/10.1088/2632-2153/acefa7.

D *** (anonymous), 2015. Convolutional networks on graphs for learning molecular fingerprints. Advances in neural information processing systems.

D *** (anonymous), 2020. A generalization of transformer networks to graphs. *arXiv preprint* arXiv:2012.09699.

F *** (anonymous), 2020. Se (3)-transformers: 3d roto-translation equivariant attention networks. Advances in neural information processing systems.

G *** (anonymous), 2017, July. Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine Learning, 70:1263-1272.

G *** (anonymous), 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* 4(2): 268-276. doi.org/10.1021/acscentsci.7b00572.

H *** (anonymous), 2017. Conformation generation: the state of the art. *Journal of chemical information and modeling*, 57(8): 1747-1756. doi.org/10.1021/acs.jcim.7b00221.

H *** (anonymous), 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840-6851.

X *** (anonymous), 2018. How powerful are graph neural networks?. *arXiv preprint* arXiv:1810.00826.

H *** (anonymous), 2019. Strategies for pre-training graph neural networks. *arXiv preprint* arXiv:1905.12265.

K *** (anonymous), 2016. A real-world perspective on molecular design: Miniperspective. *Journal of medicinal chemistry*, 59(9): 4087-4102. doi.org/10.1021/acs.jmedchem.5b01875.

L *** (anonymous), 2022. Molecular geometry pretraining with se (3)-invariant denoising distance matching. *arXiv preprint* arXiv:2206.13602.

L *** (anonymous), 2021. Pre-training molecular graph representation with 3d geometry. *arXiv preprint* arXiv:2110.07728.

S *** (anonymous), 2003. Exploring potential energy surfaces for chemical reactions: an overview of some practical methods. *Journal of computational chemistry*, 24(12): 1514-1527. doi.org/10.1002/jcc.10231.

S *** (anonymous), 2017. Quantum-chemical insights from deep tensor neural networks. *Nature communications* 8(1): 13890. doi.org/10.1038/ncomms13890.

S *** (anonymous), 2018. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24). doi.org/10.1063/1.5019779.

S *** (anonymous), 2022. 3d infomax improves gnns for molecular property prediction. In Proceedings of the 39th International Conference on Machine Learning, 162:20479-20502.

S *** (anonymous), 2020. A deep learning approach to antibiotic discovery. *Cell* 180(4): 688-702. doi.org/10.1016/j.cell.2020.01.021.

S *** (anonymous), 2017. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of cheminformatics*, 9:1-9. doi.org/10.1186/s13321-017-0222-2.

V *** (anonymous), 2017. Attention is all you need. Advances in neural information processing systems, 30.

V *** (anonymous), 2011. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661-1674. doi.org/10.1162/NECO_a_00142.

W *** (anonymous), 2021. Machine learning of free energies in chemical compound space using ensemble representations: Reaching experimental uncertainty for solvation. *The Journal of Chemical Physics*, 154(13): doi.org/10.1063/5.0041548.

W *** (anonymous), 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, *9*(2): 513-530. DOI:10.1039/C7SC02664A.

X *** (anonymous), 2022. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint* arXiv:2203.02923.

Y *** (anonymous), 2020. Graph contrastive learning with augmentations. Advances in neural information processing systems, 33:5812-5823.

Z *** (anonymous), 2022. Pre-training via denoising for molecular property prediction. *arXiv preprint* arXiv:2206.00133.

Z *** (anonymous), 2021. QSAR modeling based on conformation ensembles using a multi-instance learning approach. *Journal of Chemical Information and Modeling*, 61(10):4913-4923. doi.org/10.1021/acs.jcim.1c00692

Z *** (anonymous), 2018. Learning atoms for materials discovery. *Proceedings of the National Academy of Sciences* 115(28): E6411-E6417. doi.org/10.1073/pnas.1801181115.