Scale-Aware Graph-Based Machine Learning for Accurate Molecular Property Prediction

Gyoung S. Na, Hyun Woo Kim, and Hyunju Chang
Korea Research Institute of Chemical Technology (KRICT)
Republic of Korea
{ngs0, ahwk, hjchang}@krict.re.kr

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377905

***Abstract*—With great growth in the volume of chemical databases, machine learning receives significant attention from various scientific communities for efficient high-throughput screening of molecular properties and drug discovery on the millions of chemical compounds. In particular, graph neural networks (GNNs) have been widely studied in chemistry-related fields because a molecule is natively represented as a mathematical graph. In GNNs for the graph-level analysis, a global operation called readout is applied after node embedding to generate a graph-level embedding that represents characteristics of the whole graph. However, commonly used readouts frequently distort scale information of the graph and consequently degrade the prediction accuracy of GNNs. This problem becomes more serious in molecular machine learning because molecules have many important scale information (e.g., molecular weight and total energy). In this paper, we investigate this scale distortion problem in GNNs caused by the readouts for the first time and propose an efficient solution with a new attention-based readout. In the experiments, the proposed readout outperformed commonly used readouts on various GNN architectures.**

***Index Terms*—Deep learning, Readout, Cheminformatics**

I. Introduction

Predicting molecular properties is a key task in various scientific applications such as materials and drug discovery [1]–[3]. With great growth in the volume of chemical databases, machine learning receives significant attention to efficiently predict molecular properties based on a data-driven approach. In particular, graph neural networks (GNNs) have been widely studied in chemistry-related fields because a molecule is natively represented as a mathematical graph [4], [5]. Based on this graph representation of the molecules, GNNs achieved state-of-the-art performance in various scientific applications such as molecular property prediction [4], molecular generation [2], and materials discovery [1]. In addition to the chemistry-related fields, GNNs have been widely applied to chemical databases in computer science [1], [2], [5]–[10] because predicting molecular properties is one of the most appealing tasks of GNNs and closely related to understanding principles of natural science or discovering novel drugs and materials [1], [2], [11].

In common settings of node classification and link prediction, GNNs consist of neighborhood aggregation and fully-connected layers. In the (neighborhood) aggregation layer, the input graph-structured data in the non-Euclidean space is converted into latent node embeddings in the Euclidean space based on various aggregation schemes [12]–[14]. Then, the target values of each node are predicted through the next fully-connected layers. However, in the graph classification and regression, an additional operation called readout [15] should be implemented between the aggregation and fully-connected layers to extract a graph-level embedding based on the latent node embeddings. After calculating the graph embedding through the readout, the target values of the graph are similarly predicted by feeding the graph embedding to the fully-connected layers.

In the implementation of the readouts, $mean$ and max operations are most commonly used [15]–[20] because they are intuitive, simple, and efficient. These $mean$ and max readouts generate the graph embedding from the latent node embeddings by averaging them or extracting maximum values from each latent node embedding. In addition to these simple methods, sophisticatedly-designed readout proposed based on recurrent neural network (RNN) and has shown better prediction accuracy than the simple readouts on several datasets [21]. Although the readout is important and widely studied, it has not been studied whether it can accurately preserve graph-level attributes. However, unfortunately, commonly used readouts cannot correctly preserve the scale information of the graph (e.g., sum of degrees and molecular weight) and frequently cause significant performance degradation in prediction accuracy. We call this problem *scale distortion problem* and comprehensively investigate it for the first time.

We categorize the readouts into normalized and unnormalized functions to clarify the reasons for the scale distortion problem caused by the readouts. According to our investigation, most of the commonly used readouts are normalized. For example, the $mean$ and max readouts are normalized in row-wise and column-wise, respectively, when each row in the matrix of the latent node embeddings indicates the feature vector of one node. In addition to them, commonly used RNN-based readout is also normalized by the softmax function [21]. The fundamental characteristic of the normalized readouts is that the graph embeddings generated by them exist within the node-feature space because the graph embeddings are calculated based on the weighted sum of the node-features. However, the scale information of the graph frequently exists outside of the node-feature space. For example, molecular weight is calculated by the sum of the atomic weight of all atoms (nodes) in the molecule, and it is clearly outside of the node-feature space. In addition to the molecular weight, there

are many important scale information of the graph such as sum of degrees, total energy of the graph, and mass of the graph. In particular, these scale information of the graph is crucial in the graph-structured data of the molecular structures because the scale-related features of the molecules (e.g., molecular weight or total energy) is directly or probabilistically related to the target molecular properties in most cases [22]–[24]. Nevertheless, the normalized readouts cannot correctly extract these scale-related features of the graph from the latent node embeddings, since they are only able to generate the graph embeddings within the node-feature space.

Based on our investigation, we suggest using the unnormalized readouts to overcome the scale distortion problem. We will analytically and empirically show that the scale distortion problem can be significantly alleviated by employing the unnormalized readouts in GNNs. Also, we develop a new unnormalized readout based on the self-attention mechanism [25] for effective graph regression and classification as well as solving the scale distortion problem. In the experiments on several benchmark molecular datasets, the unnormalized readouts significantly improved the prediction accuracy of molecular properties, and our unnormalized self-attention readout achieved state-of-the-art performance among existing readouts on various GNN architectures.

II. Related work

A. Representation Capability of GNNs

The representation capability of GNNs has been widely studied from various perspectives. For the popular graph convolutional network (GCN) [12], [26] studied the representation capability of GCN based on understanding the random process from which a graph is produced. In particular, by applying a novel concept called graph moments and the modular design for GCN to graph generation, [26] improved the representation capability of GCN for graph-level analysis. However, the analysis and the empirical results of [26] is limited to GCN. Graph attention network (GAT) is a GNN using self-attention mechanism [25] in the node aggregation [13]. By exploiting the self-attention mechanism between the nodes, GAT achieved performance improvement over GCN on several benchmark datasets. Also, the representation capability of GNNs was investigated in terms of the injectivity of the aggregation scheme [14]. The authors of [14] theoretically studied the injectivity of the aggregation layer of GNNs and proposed graph isomorphism network (GIN) by enhancing the injectivity of the aggregation layer. In addition to them, the representation capability of GNNs has been widely studied in [27]–[29], but none of the existing literature investigated the scale distortion problem caused by the readout. In the experiments, we achieved further improvement by applying our unnormalized readout to the advanced GNNs such as GIN.

B. Pooling Methods and Readouts

In graph-based machine learning, the readouts are sometimes confused with the pooling methods. The fundamental role of the readout is to generate the graph-level embeddings from the latent node embeddings [15], but the role of the pooling method is to select important nodes in the node embedding process of the aggregation layers [17], [20]. For this reason, the readouts are additionally applied after the pooling layers in GNNs [16], [17], [20]. For the implementation of the readouts, $mean$, max, and their combination are commonly used [15]. The $mean$ readout generates the graph embedding by averaging the latent node embeddings, and the max readout generates the graph embedding by selecting the maximum values of each node feature from the latent node embeddings. In addition to them, an RNN-based [21] readout is also widely used to capture more complex characteristics of the graph from the latent node embeddings. In the RNN-based readout, the graph embedding is generated by the content-based attention from long short-term memory (LSTM) [30].

III. Proposed Method

In graph-based molecular machine learning, a molecule is represented as a molecular graph $G = (\mathcal{V}, A, X, E)$, where $\mathcal{V}$ is a set of atoms; $A \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ is an adjacency matrix representing the connectivities between the atoms; $X \in \mathbb{R}^{|\mathcal{V}|\times d}$ is an input node-feature matrix including d atomic features for each atom; and E is an edge-feature matrix [1], [6]. For a given labeled molecular dataset $D = \{(G_1, y_1), ..., (G_{|D|}, y_{|D|})\}$, the goal of graph-based molecular property prediction is to build a prediction model $f : G \rightarrow y$ that predicts the target property y corresponding to the input molecular graph G. Before explaining our investigation of the scale distortion problem in GNNs and the proposed solution, we define scale information as:

Definition 1 (Scale information). *Graph-level information that is proportional or inversely proportional to the number of nodes in the graph.*

A. Normalized and Unnormalized Readouts

Before investigating the scale distortion problem and its solution, first, we categorize the readouts into normalized and unnormalized as it can provide an insight for the scale distortion problem. In GNNs, the $mean$ and max readouts are most commonly used in various applications [15]–[20] due to their efficiency and simplicity. In addition to them, LSTM-based readout is also proposed to improve the prediction accuracy of GNNs [21]. A common feature of them is that they are normalized. For example, the $mean$ and LSTM-based readouts are essentially the weighted sum with the normalized weight for each node. Also, the max readout is normalized feature-wise as it selects the maximum value of each node feature in all latent node embeddings. Thus, the graph embeddings generated by the normalized readouts exist between the minimum and maximum values of the node features. However, scale information of the graph may not be in the range of node-features because it is calculated by the sum of node features in many cases (e.g., sum of degrees and molecular weight). A naive solution to make the normalized readouts can represent the scale information of the graph is to build more deep architecture so that the latent

node embeddings include the whole topological attributes of the graph. However, this solution is infeasible due to the optimization issues in the deep architecture of GNNs [15] and the node-feature convergence problem in graph convolution-based GNNs [31].

In contrast, the unnormalized readouts can generate the graph embeddings in the outside of the node-feature space because they are based on the weighted sum with the unnormalized weights. The most typical unnormalized readout is sum-based readout. The sum readout generates the graph embedding by adding latent node embedding of all nodes in the graph, so the generated graph embedding can exist outside of the node-feature space. Although the sum readout can generate the graph embedding over the node-feature space, it is hardly used in GNNs because it may cause severe fluctuation in calculating the graph embeddings and consequently makes the training unstable if the dataset contains the graphs of various scales.

B. Scale Distortion Problem

In this section, we will show the problem that the scale information of the graph can be distorted by the normalized readouts in GNNs. We call this problem *scale distortion problem*. As an example of the scale distortion problem, Fig. 1 shows two molecular graphs (G_1 and G_2) that will be represented as the same graph embedding in GNNs with the normalized readouts. Note that carbon rings are the most typical structure in chemistry.

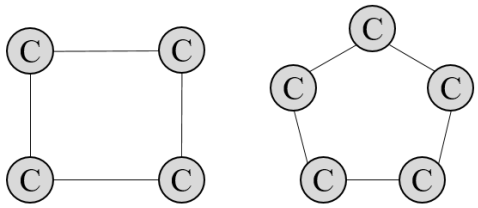


Fig. 1: Two molecular graphs of four carbon molecule (G_1) and five carbon molecule (G_2) that have the same graph embedding in GNNs with the normalized readouts.

Since two molecules have the same local structures that consist of two carbon neighbors as shown in Fig. 1, each node of them is embedded to the same latent node-features in the aggregation layers. For example, the j^{th} latent node-feature of the i^{th} node is calculated as $H_{ij} = \sigma(\sum_{q=1}^{d} X_{iq} W_{qj})$ in the graph convolution [12], where σ is a nonlinear activation and W is a trainable weight matrix of the graph convolution layer. If each row of the input node-feature matrix H is the same each other like the carbon rings, $\sum_{q=1}^{d} X_{iq} W_{qj}$ for all i will be the same. For the latent node embeddings of the two carbon rings, these carbon rings are represented as the completely same vector when $mean$, max, and other normalized readouts are applied. Thus, the scale information of the graph is distorted.

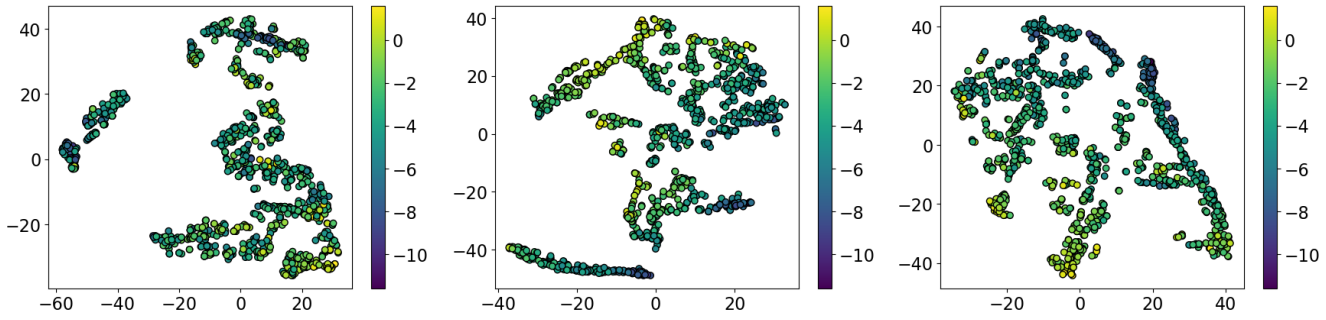
In addition to the above simple example, the scale distortion problem can occur even if the molecules are different in the whole topology because their scale information can still be distorted if they have similar local structures with different scales. In particular, the scale distortion problem becomes more common and serious in the molecular datasets due to the three characteristics of them:

- Most of the molecular datasets contain topologically similar molecules because the physical and chemical experiments are conducted with a focus on a set of several base molecules due to cost and time limitations in the experiment [22]–[24], [32]–[34].
- Because molecules are generated according to certain chemical rules, the molecules frequently share similar local structures such as carbon rings in Fig. 1.
- In many cases, molecules have different properties even though they are topologically similar, and it is sometimes related to the scale information of them (e.g., aqueous solubility).

To empirically show the scale distortion problem, we drew the distributions of 1,128 organic molecules in the estimated solubility (ESOL) dataset [22] for their aqueous solubilities. Note that aqueous solubility is a typical molecular property that is roughly correlated to the scale of the molecules. We used a randomly-initialized GCN to project the graph-structured data into Euclidean space while preserving their density based on structural similarity. Then, we extracted the graph embeddings for each molecule using the randomly-initialized GCN and visualized them using t-SNE [35].

Fig. 2 shows the distributions of the molecules based on their graph embeddings calculated by GCNs with three readouts: (a) $mean$ readout (normalized readout). (b) concatenation of $mean$ readout and molecular weight of the molecules. (c) sum readout (unnormalized readout). In the figure, each point is a molecule, and the color of the points indicates the value of the aqueous solubility. As shown in Fig. 2-(a), the distribution of the molecules is natively mixed to their aqueous solubilities. Moreover, some molecules have completely different solubilities even though they are structurally similar. After simply concatenating one of the scale information called molecular weight to graph embeddings, the distribution became more discriminative for the solubilities as shown in Fig. 2-(b). This performance improvement in the graph embedding is trivial because aqueous solubility is roughly correlated to the scale of molecules, and the scale information of the molecules is directly provided by concatenating the molecular weight to the graph embeddings. However, as shown in Fig. 2-(c), although we did not provide any scale information of the molecules in the graph embedding process, we were able to achieve the embedding results similar to Fig. 2-(b) by applying sum readout which is a simple unnormalized readout. This example clearly shows that scale information of the graphs is crucial for accurate prediction in molecular machine learning, and it can be efficiently extracted by unnormalized readouts.

In the previous work of GNNs, we can also observe the scale distortion problem. In the experiments of [25], the classification accuracies on molecular datasets were dropped in



(a) $mean$ readout (b) $mean$ readout with weight (c) sum readout

Fig. 2: Distributions of the molecules for aqueous solubility in the randomly-projected space. X and Y axes indicates 2-dimensional features generated by t-SNE [35].

a hierarchical structure that the pooling layer and the readout are stacked on each aggregation layer. In their implementation, the concatenation of two normalized readouts was used to generate the graph embeddings. Although the reason for the performance degradation was not discussed in [25], we can clarify that the reason for the performance degradation is the scale distortion problem caused by the normalized readouts because the more normalized readouts were stacked, the more prediction accuracies of the molecular properties were dropped.

C. Unnormalized Readouts for Scale Distortion Problem

So far, we described the scale distortion problem and investigated the reason for it by categorizing the readouts into the normalized and unnormalized operations. As we mentioned before, the scale distortion problem can be solved by exploiting the unnormalized readout without additional complexity and model modifications. One of the simplest unnormalized readouts is sum-based readout. Since the sum readout generates the graph embedding by adding all latent node embeddings of the graph, it can generate the graph embeddings in the outside of the node-feature space. In section IV-A, we will experimentally show that the sum readout can effectively extract scale information from the graph, whereas two normalized $mean$ and max readouts cannot do that.

Nevertheless, the sum readout still has a limitation that it may not be able to capture the complex characteristics of the graph from the latent node embeddings because it simply generates the graph embeddings by adding all latent node embeddings. To overcome the limitation of the sum readout, we applied a well-known self-attention mechanism [25] to the readout. A typical self-attention based readout can be defined as:

$$\mathbf{z} = \sum_{n=1}^{N} \text{softmax}(h_{gate}(H_{n,:}^{(L)}); H^{(L)}) \odot H_{n,:}^{(L)}, \qquad (1)$$

where N is the number of nodes in the graph; L is the number of aggregation layers in GNNs; h_{gate} indicates a simple neural network to calculate the attention scores; $H_{n,:}$ is the n^{th} row vector of the latent node embedding matrix H; and $\odot$ is element-wise multiplication. Obviously, the attention scores in Eq. (1) are normalized by the softmax function, and consequently the whole readout is also normalized.

To solve the scale distortion problem caused by the normalized readouts, we unnormalize the self-attention based readout in Eq. (1) as:

$$\mathbf{z} = \sum_{n=1}^{N} \sigma(h_{gate}(H_{n,:}^{(L)})) \odot H_{n,:}^{(L)}, \qquad (2)$$

where σ is an unnormalized function in node-wise, such as hyperbolic tangent and softplus [40]. By simply converting the softmax function to the unnormalized function, the whole self-attention based readout is efficiently unnormalized. In the experiment section, we will comprehensively evaluate the effectiveness of our unnormalized attention readout in Eq. (2) on benchmark molecular datasets containing various target properties that are related to or not related to the scale information of the molecules.

IV. Experiment

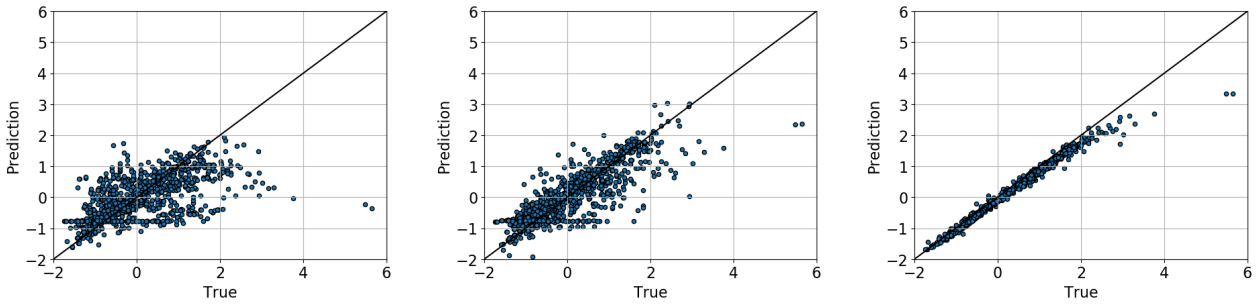
We conducted comprehensive experiments on molecular datasets to validate the effectiveness of our analysis and the proposed readout for the scale distortion problem. As a base model of GNNs, we used graph convolutional network (GCN) [12], graph attention network (GAT) [13], and graph isomorphism network (GIN) [14]. For these base GNNs, we applied four normalized and two unnormalized readouts in predicting molecular properties. To compare the prediction capabilities of the readouts, we generated six GNN-based models by only changing the readouts: (1) GNN-mean; (2) GNN-max; (3) GNN-attn; (4) GNN-LSTM; (5) GNN-sum; (6) GNN-uattn. Note that the suffix of the generated GNNs indicates the type of their readout. We denoted the normalized and unnormalized attention readouts by attn and uattn, respectively.

All GNNs and experiment scripts were implemented on the PyTorch framework[1] and the PyTorch-Geometric library[2]. We used a well-known chemical library called RDKit[3] to

[1]https://pytorch.org/
[2]https://pytorch-geometric.readthedocs.io/en/latest/
[3]https://www.rdkit.org/

TABLE I: Characteristics of the benchmark graph-based molecular datasets used in the experiments.

Dataset	Target property	Unit	Target range	Average of targets	Task	# of molecules
ESOL [22]	Aqueous solubility	log(mol/L)	[-11.60, 1.58]	-3.05 ± 2.10	Regression	1,128
Freesolv [32]	Hydration free energy	kcal/mol	[-25.57, 3.43]	-3.80 ± 3.84	Regression	642
Lipophilicity [36]	Water distribution coefficient	log$D$	[-1.50, 4.50]	2.20 ± 1.20	Regression	4,200
PDBbind [37]	Binding affinity	kcal/mol	[0.4, 9.3]	6.20 ± 1.70	Regression	9,880
QM7 [33], [34]	Atomization energy	kcal/mol	[-2.90, 5.11]	-1.41 ± 0.99	Regression	6,830
QM9-IP [23], [24]	Isotropic polarizability	bohr^3	[6.31, 196.62]	75.19 ± 8.19	Regression	133,885
QM9-HLG [23], [24]	HOMO-LUMO gap	Hartree	[0.02, 0.62]	0.25 ± 0.05	Regression	133,885
BACE [38]	Binding results for inhibitors	-	$\{0, 1\}$	-	Binary classification	1,513
BBBP [39]	Blood-brain barrier penetration	-	$\{0, 1\}$	-	Binary classification	2,039



(a) GCN with $mean$ readout (b) GCN with max readout (c) GCN with sum readout

Fig. 3: Experimental results of the molecular weight regression. X and Y axes indicate the true molecular weight and the predicted molecular weight, respectively. Note that $mean$ and max readouts are normalized, and sum readout is unnormalized.

convert the molecular structures into the molecular graphs. All experiments of this paper were conducted in a machine of Intel CPU i7-9700K 3.60GHz with 64GB RAM and NVIDIA RTX 2080 Ti.

A. Molecular Weight Regression

Before the experimental evaluations on real-world molecular datasets, we conducted a simple regression experiment that predicts the molecular weight from the molecular graph. Since molecular weight is the most fundamental scale information that can be directly and simply calculated from the molecular structure, we can evaluate how well each readout extracts the scale information from the molecular structures through this experiment. For the regression, we calculated the molecular weights of the molecules in the ESOL dataset using RDKit, and the calculated molecular weights were used as a target property. Since the molecular weight is defined as the sum of the atomic weights, we included the atomic weight in the input node-features. Chemically, this molecular weight regression is a straightforward problem because the atomic weights are given, and the molecular weight is just the sum of the given atomic weights.

Fig. 3 presents the regression results of GCN-mean, GCN-max, and GCN-sum. In the figures, X and Y axes indicate the true and predicted target values, respectively. Hence, the closer each point is to the $y = x$ line, the higher the prediction accuracy. As shown in Fig. 3-(a) and (b), the GCNs with the normalized readouts were unable to accurately learn the relationship between the input atomic weights and the molecular weight even though their relationship is extremely simple. In contrast, the GCN with the unnormalized readout effectively extracted the molecular weights of the molecules from the input-node features as shown in Fig. 3-(c). These experimental results clearly show that the unnormalized readouts are more effective than the commonly used normalized readouts in extracting the scale information of the graphs.

B. Molecular Property Prediction and Classification

Table II shows the prediction error and the classification accuracy of the GCNs with the normalized and unnormalized readouts on molecular datasets. The prediction error and the classification accuracy were measured by root mean square error (RMSE) and F1-score, respectively. On the bottom of the table, the average rank of the prediction accuracies for each readout is reported. In the experiments, the GCNs with the unnormalized readouts (GCN-sum and GCN-uattn) outperformed all GCNs with the normalized readouts even though sum readout (GCN-sum) is extremely simpler than attention and LSTM based readouts (GCN-attn and GCN-LSTM), as shown in Table II. These experimental results clearly show that preserving the scale information of the graph is crucial to accurately predict molecular properties in many cases. In particular, the prediction errors of GCN-uattn were significantly reduced by 45.79% and 42.07% on

TABLE II: Regression and classification results of the GCNs with the normalized and unnormalized readouts on the benchmark datasets. RMSE and F1-score are used to measure the prediction error and the classification accuracy, respectively. The best performance is marked as the bold font, and the second best performance is denoted by the underline.

Algorithm		ESOL	FreeSolv	Lipophilicity	PDBbind	QM7	QM9-IP	QM9-HLG	BACE	BBBP	Avg. rank
Normalized readout	GCN-mean	0.668	0.576	0.696	0.822	0.792	0.416	0.722	0.808	0.867	4.56
	GCN-max	0.537	0.462	0.717	0.841	0.772	0.409	0.732	0.770	0.867	4.78
	GCN-attn	0.651	0.568	0.700	0.822	0.785	0.392	0.720	0.792	0.870	4.11
	GCN-LSTM	0.579	0.449	0.649	0.838	0.771	0.294	0.755	0.773	0.846	4.00
Unnormalized readout	GCN-sum	**0.356**	0.547	0.678	0.822	0.711	0.249	0.721	0.805	0.873	2.44
	GCN-uattn	0.362	**0.376**	**0.622**	**0.805**	**0.693**	**0.241**	**0.712**	**0.815**	**0.881**	1.11

TABLE III: Regression and classification results of the GATs with the normalized and unnormalized readouts.

Algorithm		ESOL	FreeSolv	Lipophilicity	PDBbind	QM7	QM9-IP	QM9-HLG	BACE	BBBP	Avg. rank
Normalized readout	GAT-mean	0.738	0.506	0.747	0.844	0.829	0.547	0.759	0.763	0.868	4.67
	GAT-max	0.730	0.437	0.836	0.861	0.868	0.648	0.774	0.770	0.860	5.22
	GAT-attn	0.746	0.496	0.766	0.841	0.819	0.468	0.755	0.776	0.866	4.11
	GAT-LSTM	0.694	0.453	**0.733**	0.857	0.808	0.418	**0.750**	0.753	0.862	3.44
Unnormalized readout	GAT-sum	0.385	0.368	0.746	0.792	0.716	0.366	0.752	0.771	0.880	2.22
	GAT-uattn	**0.361**	**0.357**	0.744	**0.791**	**0.704**	**0.357**	0.752	**0.784**	**0.882**	1.33

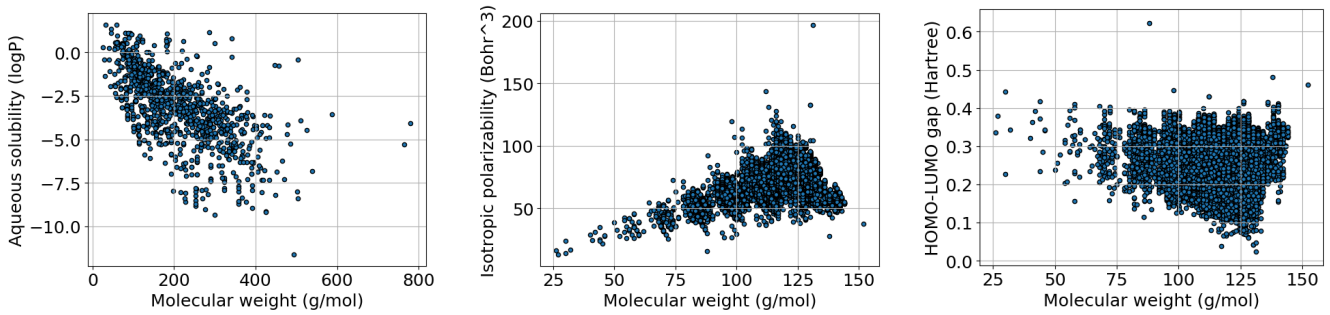
TABLE IV: Regression and classification results of the GINs with the normalized and unnormalized readouts.

Algorithm		ESOL	FreeSolv	Lipophilicity	PDBbind	QM7	QM9-IP	QM9-HLG	BACE	BBBP	Avg. rank
Normalized readout	GIN-mean	0.461	0.357	0.743	0.824	0.762	0.391	0.717	0.794	0.872	4.00
	GIN-max	0.477	0.365	0.726	0.827	0.781	0.439	0.742	0.786	0.864	5.33
	GIN-attn	0.456	0.360	0.654	0.812	0.766	0.389	0.713	0.798	0.868	3.22
	GIN-LSTM	0.457	0.360	0.686	0.841	0.797	0.278	0.777	0.770	0.850	5.11
Unnormalized readout	GIN-sum	0.365	0.342	0.664	**0.795**	0.703	0.245	0.714	0.793	0.886	2.22
	GIN-uattn	**0.345**	**0.322**	**0.629**	0.798	**0.683**	**0.232**	**0.709**	**0.801**	**0.887**	1.11

the ESOL and QM9-IP datasets compared to the commonly used GCN-mean, respectively. The significant improvement in predicting aqueous solubility on the ESOL dataset is important in chemistry and biology because aqueous solubility is one of the essential and important properties to determine the applicability of drugs and vaccines for humans [41]–[44]. In addition to GCN, we measured the performance improvement by the unnormalized readouts in the currently proposed GAT and GIN. The performance improvement by the unnormalized readouts was consistent in both GAT and GIN architectures as shown in Table III and IV. Further performance improvement in GIN is interesting because even state-of-the-art GNN can still not capture the scale information of the graphs in the graph embedding process.

The improvement in the prediction accuracy on the ESOL and QM9-IP datasets can be rationalized by recent scientific observations. The aqueous solubility has been generally expected to be closely related to the size of the molecules, and these expected relationships were also systemically revealed in [45]. For the isotropic polarizability, although it does not have a clear relationship with the size of the molecules, their relationship can be described probabilistically. The isotropic polarizability in the organic molecules is closely related to the existence of ring structure in the molecules [46], and the probability for the existence of such rings naturally increases the size of molecules. Thus, the significant performance improvements on the ESOL and QM9-IP datasets are scientifically reasonable. Furthermore, the marginal performance improvement on the QM9-HLG dataset is also rationalized scientifically because its target property (HOMO-LUMO gap) is defined as the energy difference between two specific portions of the molecule [47]. In other words, HOMO-LUMO gap of the molecules is more relevant to the local structures of the molecules than to the global structures or size of the entire molecules in this dataset.

To statistically rationalize the performance improvement by the unnormalized readouts, we draw the distribution of the target properties for the molecular weight that is one of the most essential scale information in the molecular systems. Fig. 4 shows the distributions of aqueous solubility,



(a) Aqueous solubility (b) Isotropic polarizability (c) HOMO-LUMO gap

Fig. 4: Distributions of aqueous solubility, isotropic polarizability, and HOMO-LUMO gap for the molecular weight. X and Y axes are the molecular weight (g/mol) and the target property, respectively.

isotropic polarizability, and HOMO-LUMO gap for the molecular weight. Aqueous solubility and isotropic polarizability of the molecules have relatively strong correlations with the molecular scale (molecular weight), and their correlation coefficients were -0.6387 and 0.3829, respectively. Thus, the GCNs with the unnormalized readouts were able to improve the prediction accuracy because the unnormalized readouts could extract the scale information from the latent node embeddings. In contrast, HOMO-LUMO gap showed a weak relationship with the molecular weight, and its correlation coefficient was -0.0074. For this reason, the performance improvement of the GCNs with the unnormalized readouts was marginal.

V. Conclusions

This paper investigated the scale distortion problem of GNNs in the graph-level analysis for the first time and demonstrated that it is crucial for accurate prediction of the graph-level target values in various scientific applications of chemistry. To solve the scale distortion problem efficiently, we suggested to use the unnormalized readouts and proposed a new unnormalized readout based on the self-attention mechanism. In the experiments on extensive benchmark datasets, the unnormalized readouts outperformed existing readouts in predicting molecular properties on the several molecular datasets where the scale information of the graph is important. In particular, the unnormalized readouts remarkably improved the prediction accuracy of the aqueous solubility of the molecules, which is the most fundamental property of drugs and vaccines for humans. Furthermore, our unnormalized attention readout not only achieved the best prediction accuracy on molecular datasets but also showed comparable accuracy with the most popular readout on general molecular datasets.

Acknowledgment

This research was supported by the core KRICT project from the Korea Research Institute of Chemical Technology (SI2051-10).

References

[1] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accuracy and interpretable prediction of material properties," *Physical Review Letters*, vol. 120, p. 145301, 2018.

[2] J. You, B. Liu, R. Ying, V. Pande, and J. Leskovec, "Graph convolutional policy network for goal-directed molecular graph generation," *Conference on Neural Information Processing Systems (NIPS)*, 2018.

[3] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Comput. Mater.*, vol. 5, Aug 2019.

[4] Z. Wu, B. Ramsundar, E. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: a benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, pp. 513–530, 2018.

[5] C. Lu, Q. Liu, C. Wang, Z. Huang, P. Lin, and L. He, "Molecular property prediction: A multilevel quantum interactions modeling perspective," *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[6] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: a benchmark for molecular machine learning," *Chemical Science*, vol. 9, pp. 513–530, 2018.

[7] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, "A graph-convolutional neural network model for the prediction of chemical reactivity," *Chemical Science*, vol. 10, pp. 370–377, 2019.

[8] W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," *International Conference on Machine Learning (ICML)*, 2018.

[9] B. Samanta, A. DE, G. Jana, P. K. Chattaraj, N. Ganguly, and M. G. Rodriguez, "Nevae: A deep generative model for molecular graphs," *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[10] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *International Conference on Machine Learning (ICML)*, 2017.

[11] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature 559*, pp. 574–555, 2018.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *International Conference on Learning Representations (ICLR)*, 2017.

[13] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *International Conference on Learning Representations (ICLR)*, 2018.

[14] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *International Conference on Learning Representations (ICLR)*, 2019.

[15] Z. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *arXiv:1812.08434*, 2018.

[16] J. B. Lee, R. Rossi, and X. Kong, "Graph classification using structural attention," *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

[17] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," *Conference on Neural Information Processing Systems (NIPS)*, 2018.
[18] K. Xu, C. Li, Y. Tian, T. Sonobe, K. ichi Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," *International Conference on Machine Learning (ICML)*, 2018.
[19] C. Cangea, P. Velickovic, N. Jovanovic, T. Kipf, and P. Lio, "Towards sparse hierarchical graph classifiers," *arXiv preprint arXiv:1811.01287*, 2018.
[20] J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," *International Conference on Machine Learning (ICML)*, 2019.
[21] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," *International Conference on Learning Representations (ICLR)*, 2016.
[22] J. S. Delaney, "Esol: Estimating aqueous solubility directly from molecular structure," *Chemical Information and Computer Sciences*, vol. 44, pp. 1000–1005, 2004.
[23] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structured and properties of 134 kilo molecules," *Nature Scientific Data*, vol. 1, 2014.
[24] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17," *Physical of Chemical Information and Modeling*, vol. 52, pp. 2864–2875, 2012.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Conference on Neural Information Processing Systems (NIPS)*, 2017.
[26] N. Dehmamy, A.-L. Barabasi, and R. Yu, "Understanding the representation power of graph neural networks in learning graph topology," *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
[27] F. Scarselli, M. Gori, A. C. Tsoi, M. Hangenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, 2008.
[28] R. L. Murphy, B. Srinivasan, V. Rao, and B. Ribeiro, "Relational pooling for graph representations," *International Conference on Machine Learning (ICML)*, 2019.
[29] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and leman go neural: Higher-order graph neural networks," *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.
[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computations)*, vol. 9 (8), pp. 1735–1780, 2018.
[31] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.
[32] D. L. Mobley and J. P. Guthrie, "Freesolv: A database of experimental and calculated hydration free energies, with input files," *Journal of computer-aided molecular design*, vol. 28, pp. 711–720, 2014.
[33] L. C. Blum and J.-L. Reymond, "970 million druglike small molecules for virtual screening in the chemical universe database gdb-13," *Journal of the American Chemical Society*, vol. 131, pp. 8732–8733, 2009.
[34] M. Rupp, A. Tkatchenko, K.-R. Muller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical Review Letters*, vol. 108, p. 058301, 2012.
[35] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[36] M. Wenlock and N. Tomkinson, "Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds," *https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL3301361/*, 2015.
[37] R. Wang, X. Fang, Y. Lu, and S. Wang, "The pdbbind dataset: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures," *Journal of Medicinal Chemistry*, vol. 47, pp. 2977–2980, 2004.
[38] G. Subramanian, B. Ramsundar, V. Pande, and R. A. Denny, "Computational modeling of beta-secretase 1 (bace-1) inhibitors using ligand based approaches," *J. Chem. Inf. Model.*, vol. 56, pp. 1936–1949, 2016.
[39] I. F. Martins, A. L. Teixeira, L. Pinheiro, and A. O. Falcao, "A batesian approach in silico blood-brain barrier penetration modeling," *J. Chem. Inf. Model.*, vol. 52, pp. 1686–1697, 2012.
[40] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *International Conference on Atrificial Intelligence and Statistics (AISTATS)*, 2011.
[41] T. Heimbach, D. Fleisher, and A. Kaddoumi, "Overcoming poor aqueous solubility of drugs for oral delivery," *Prodrugs: Challenges and Rewardss*, pp. 157–215, 2007.
[42] M. Ishikawa and Y. Hashimoto, "Improvement in aqueous solubility in small molecule drug discovery programs by disruption of molecular planarity and symmetry," *J. Med. Chem*, vol. 54, pp. 1539–1554, 2011.
[43] H. Chen, C. Khemtong, X. Yang, X. Chang, and J. Gao, "Nanonization strategies for poorly water-soluble drugs," *Drug Discovery Today*, vol. 16, pp. 354–360, 2011.
[44] B. E. Rabinow, "Nanosuspensions in drug delivery," *Nat Rev Drug Discov*, vol. 3, pp. 785–796, 2004.
[45] J. Tolls, J. van Dijk, E. J. M. Verbruggen, J. L. M. Hermens, B. Loeprecht, and G. Schuurmann, "Aqueous solubility-molecular size relationships: A mechanistic case study using c10-toc19-alkanes," *J. Phys. Chme. A*, vol. 106 (11), pp. 2760–2765, 2002.
[46] C. S. Ewig, M. Waldman, and J. R. Maple, "Ab initio atomic polarizability tensors for organic molecules," *J. Phys. Chme. A*, vol. 106 (2), pp. 326–334, 2002.
[47] J. ichi Aihara, "Reduced homo-lumo gap as an index of kinetic stability for polycyclic aromatic hydrocarbons," *J. Phys. Chme. A*, vol. 103, pp. 7487–7495, 1999.