Chemistry Guided Molecular Graph Transformer

Peisong Niu* Tian Zhou *† Qingsong Wen
Liang Sun Tao Yao
{niupeisong.nps, tian.zt, qingsong.wen}@alibaba-inc.com
{liang.sun, tao.yao}@alibaba-inc.com

Abstract

Classic methods to calculate molecular properties are insufficient for large amounts of data. The Transformer architecture has achieved competitive performance on graph-level prediction by introducing general graphic embedding. However, the direct spatial encoding strategy ignores important inductive bias for molecular graphs, such as aromaticity and interatomic forces. In this paper, inspired by the intrinsic properties of chemical molecules, we propose a chemistry-guided molecular graph Transformer. Specifically, motif-based spatial embedding and distance-guided multi-scale self-attention for graph Transformer are proposed to predict molecular property effectively. To evaluate the proposed methods, we have conducted experiments on two large molecular property prediction datasets, ZINC, and PCQM4M-LSC. The results show that our methods achieve superior performance compared to various state-of-the-art methods. Codes are available at `https://github.com/PSacfc/chemistry-graph-transformer`.

1 Introduction

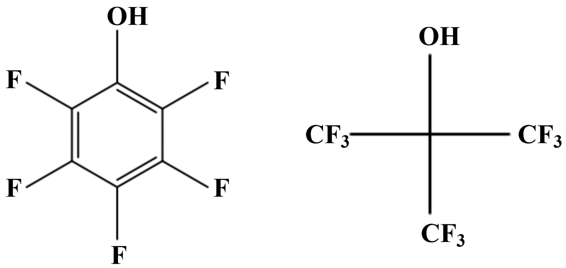
Molecular properties prediction is one of the most critical tasks and plays a vital role in many downstream applications, from material discovery to drug design. Classic techniques, such as Density Functional Theory (DFT), are computationally too expensive to deal with large amounts of data. As molecules and graphs share similar structures, it is natural to model molecules as graphs for representation learning in deep neural networks. In particular, graph neural networks (GNNs) [1] have achieved considerable progress in chemistry calculation [2]. However, standard GNNs in the neighborhood aggregation paradigm suffer from limited discriminative power in distinguishing high-order graph structures as opposed to some low-order motif. Moreover, when GNNs go deeper, the over-smoothing [3] problem gets worse for the message passing paradigm.

Transformer [4] has achieved remarkable progress in multiple domains [5, 6, 7] and could become a new powerful workhorse for graph representation learning. Recently, Graphormer [8] added a set of graph spatial-related encodings into the vanilla Transformer and achieved impressive performance on graph-level prediction tasks. EGT [9] added a dedicated pathway for pairwise structural information and proposed Edge-augmented Graph Transformer. Although Graphormer has a clear advantage of generality, spatial encoding strategy ignores some most important molecular properties that commonly persist in nature, like aromaticity [10] and interatomic forces [11].

Aromaticity [12] is a property of cyclic molecular structure that gives highly increased stability and completely different chemical properties. Aromatic rings exhibit robust stabilization and connectivity with their delocalized pi-electron around the rings. As shown in Fig. 1, fluorine atoms (F) on C6F5OH are strongly influencing the hydroxy group (OH) through the aromatic ring's pi-electrons, causing C6F5OH to be one of the most acidic phenols with $pK_a = 5.5$, where pK_a is the acid dissociation

* Equal contribution
† Corresponding authors

NeurIPS 2022 AI for Science Workshop.



(a) C6F5OH (b) C4F9OH

Figure 1: Aromaticity and acidicity.

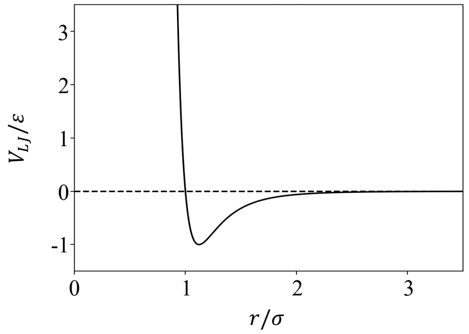


Figure 2: Lennard-Jones potential

constant. That lower value of the acid dissociation constant represents strong acid. However, for C4F9OH with more fluorine atoms, the pK_a is only 7.05 through regular bonds, and the connectivity strength is simply determined by spatial distance. Inspired by this phenomenon, we propose a motif-based shortest path distance (SPD) embedding following the inductive bias as the interaction between atoms in a single aromatic ring is vital.

Interatomic forces are composed of attraction and repulsion. Interatomic forces can also be described as potential energy. Lennard-Jones potential [13] is the simplest interatomic interaction model widely used:

$$V_{LJ}(r) = 4\varepsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}], \tag{1}$$

where r is the distance between atoms. As shown in Fig. 2, different distance scales lead to different kinds of forces. Two atoms repel each other at a close distance and attract each other at a further distance. Thus, motivated by multi-scale self-attention [14], we propose a distance-guided multi-scale self-attention for molecular graphs.

In this paper, we propose a chemistry-guided molecular graph Transformer. Inspired by Graphormer, we take nodes in a graph as tokens fed into Transformer and propose two components to help represent molecule graphs more naturally with chemical intuition. Firstly, we redefine distance embedding related to connectivity and propose motif-based spatial embedding to explicit model molecular aromatic. Secondly, we propose a distance-guided multi-scale self-attention. Only node pairs under specific distance scales participate in attention bias calculation for each Transformer layer. We have conducted extensive experiments on molecular property prediction datasets, including ZINC [15] and PCQM4M-LSC [16]. The results show that our methods achieve superior performance compared to state-of-the-art methods.

2 Preliminaries

Notations An undirected graph can be denoted as $G = (V, E)$ where $V = \{v_1, v_2, ..., v_n\}$ is the node set, $n = |V|$ is the number of nodes and E is the edge set. Let $X = \{x_1, x_2, ..., x_n\}$ be the node feature set where x_i is the feature associated with node v_i.

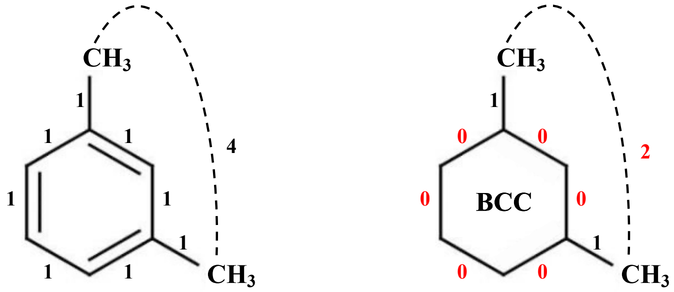
Transformers Multi-head self-attention is the main component of Transformer. Let $H \in \mathbb{R}^{n\times d}$ denote the input, where d is the hidden dimension. The self-attention module projects the input H into three matrices Q, K, V by three parameter matrices $W_Q \in \mathbb{R}^{d\times d_Q}, W_K \in \mathbb{R}^{d\times d_K}, W_V \in \mathbb{R}^{d\times d_V}$. These three matrices are projected back to each position by $W_O \in \mathbb{R}^{hd_V \times d}$ where h is the number of heads. The multi-head self-attention can be formalized as follows:

$$\mathrm{MultiHead}(H) = [head_1, head_2, ..., head_h]W^O, \tag{2}$$

$$head_i = \mathrm{softmax}(A_i)V_i, \tag{3}$$

$$A_i = \frac{QK^T}{\sqrt{d_K}}, \tag{4}$$

$$Q = HW^Q, K = HW^K, V = HW^V. \tag{5}$$



(a) Atom-based SPD

(b) Motif-based SPD

Figure 3: Atom-based SPD vs. motif-based SPD.

Apart from the self-attention module, a position-wise feed-forward network (FFN) is also one of the key components in Transformer. Let H_l be the hidden state of the l^{th} layer, H_{l+1} can be calculated as follow:

$$H_{l+1} = \mathrm{LayerNorm}(Z_l + \mathrm{FFN}(Z_l)), \tag{6}$$

$$Z_l = \mathrm{LayerNorm}(H_l + \mathrm{MultiHead}(H_l)), \tag{7}$$

where $\mathrm{LayerNorm}(\cdot)$ represents layer normalization in [17].

3 Methods

In this section, we redefine distance according to connectivity and propose motif-based spatial embedding as the bias of self-attention in each layer. Then, we propose distance-guided multi-scale self-attention in each token that pays attention to different distance scales in different layers. In addition to these methods, we utilize centrality encoding, and edge encoding in Graphormer [8] for augmenting structural information.

3.1 Motif-Based Spatial Embedding

We suppose the function distance between node pairs within a strong connectivity component is smaller than its counterpart between node pairs among unconnected components. It is guided by the vital concept of aromaticity in chemistry. In chemistry, aromaticity is a property of cyclic and planar molecular structures with pi bonds in resonance that gives stronger stability than saturated compounds and other geometric non-cyclic arrangments with the same set of atoms. The delocalized electrons will significantly impact the molecular property as a whole other than on the individual atom that consists of the molecule. Because most aromatic components in molecular graphs are rings, we consider using a bi-connected component (BCC) [18] of an undirected graph to label such inductive bias. The bi-connected component is a sub-graph in which any cut of edges has no influence on whether it is connected. We use a shared shortened distance embedding to model the effect of delocalized electrons on the individual atoms of aromatic rings. As in Fig. 3, the benzene ring is regarded as a bi-connected component. The distance between carbons of methyl groups reduces from 4 to 2.

Firstly, for graph $G(V, E)$, bi-connected components are found by [18]. Then, all nodes in a single bi-connected component are treated as a single node to construct a new acyclic graph $G^{bcc}(V^{bcc}, E^{bcc})$ and $dis^{bcc}(s, t)$ represents the SPD between v_s^{bcc} and v_t^{bcc}. Thus, for nodes $v_i \in v_s^{bcc}$ and $v_j \in v_t^{bcc}$, the motif-based SPD $dis^{motif}(i, j)$ can be calculated in Eq. 8 as follows:

$$dis^{motif}(i,j) = \left\{ \begin{array}{cc} 0, & s = t \\ dis^{bcc}(s,t), & s \neq t \end{array} \right. . \tag{8}$$

Finally, for self-attention of h^{th} head A_h, motif-based spaial embedding utilizes motif-based SPD to calculate embedding and serves as a bias term:

$$\widetilde{A}_h(i,j) = A_h(i,j) + \mathrm{emb}(dis^{motif}(i,j)). \tag{9}$$

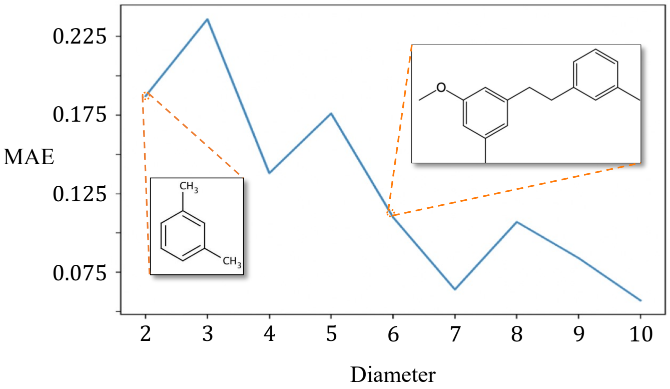


Figure 4: Structural debilitating problem.

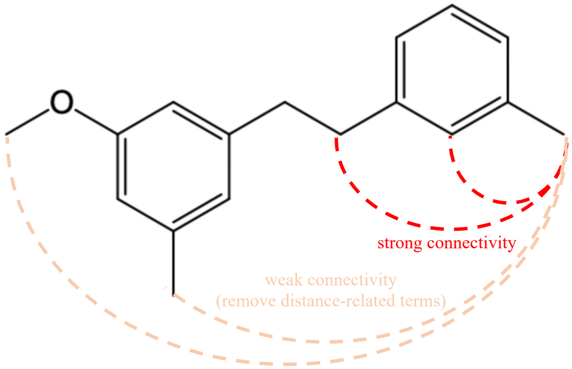


Figure 5: Distance guided multi-scale self-attention.

Structural Debilitating Problem For data with solid connectivity, structural information debilitating problem is introduced. To analyze the problem, we split ZINC (the complete details are provided in the section of experiments) by the diameter of the motif-based graph. The diameter of the graph means the maximum SPD of all node pairs. As results shown in Fig. 4, the performance in small diameter is relatively poor.

3.2 Distance Guided Multi-Scale Self-Attention

To eliminate structural debilitating and utilize interatomic forces, we introduce multi-scale self-attention. However, unlike NLP tasks, the distance between graph nodes is not regular as tokens in a sequential language signal. Thus, multi-scale self-attention [14] is not directly applicable to graph-related tasks. Distance-guided multi-scale attention is then proposed to tackle this problem, and heads in different layers work on various scales.

As shown in Fig. 5, the interaction between weakly connected pairs with farther distances is ignored. Thus, for a single head, we remove distance-related terms in the attention matrix when the distance between nodes exceeds ω^{multi}:

$$\widetilde{A}_h^{multi}(i,j) = \begin{cases} \widetilde{A}_h(i,j), & dis^{bcc}(s,t) \leq \omega^{multi} \\ A_h(i,j), & dis^{bcc}(s,t) \geq \omega^{multi} \end{cases} . \tag{10}$$

Since scale selection is essential to capture hierarchical information, we design a Fibonacci scale distribution from the intuitive graph perspective. Let $L(L \geq 2)$ denote the number of total layers and the distance scale of l_{th} layer ω_l^{multi} is computed as follow:

$$\omega_l^{multi} = S \cdot \frac{z_l}{\max(z)}, \tag{11}$$

$$z_l = \begin{cases} \alpha \cdot z_{l+1} + (2-\alpha) \cdot z_{l+2}, & l/k < L/k - 1 \\ 2, & l/k = L/k - 1 \\ 1, & l/k = L/k \end{cases}, \tag{12}$$

where $\alpha (0 \leq \alpha \leq 2)$, S and k are hyper-parameters. Here α controls the scale distribution, S controls the maximum scale size, and k represents that two adjacent k layers share the scale size.

4 Experiments

In this section, we conduct experiments on two molecular property prediction datasets, ZINC [15] and OGB-LSC [16]. Then, we provide ablation studies to demonstrate the significance of each proposed component. Finally, we analyze that distance-guided multi-scale self-attention effectively eases the debilitating structural problem.

Datasets ZINC is one of the most popular real-world molecular datasets of 250K graphs. Similar to benchmark [19], we choose the subset (12K) to regress molecular constrained solubility. In KDD Cup 2021, OGB-LSC was provided to the community to encourage the development of state-of-the-art graph ML models for sizeable molecular graph datasets. The graph regression challenge of OGB-LSC, PCQM4M-LSC, is a quantum chemistry dataset containing $3.8M$ molecular graphs. Specifically, the task is to predict the HOMO-LUMO energy gap of molecules given their 2D molecular graphs.

4.1 Graph Representation

Baseline We benchmark our model with various GNNs such as GCN [20], GAT [21], GIN [22] and MoleculeX [23]. In addition, we compare the recently proposed random walk-based method CRAWL [24]. Especially, Transformer-based model SAN [25] , GT[26] and the challenge winning solution Graphormer [8], is also compared.

Settings For both training procedures, we use Adam as the optimizer and adopt several warm-up steps followed by a linear decay learning rate scheduler. For the ZINC dataset, the large model is not encouraged. Thus parameters of the model are limited to under $500K$. And we conduct experiments with training procedures under 4 different seeds. The detailed experimental settings are shown in Table 1.

Table 1: Experimental settings.

	ZINC	PCQM4M-LSC$_{SMALL}$	PCQM4M-LSC
#Layers	12	6	12
Hidden Dimensions	80	512	768
FFN Inner Dimension	80	512	768
#Attention Heads	32	32	32
Maximum Scale Size	8	15	15
#Adjacent Scale Groups	2	1	2
Scale Distribution Weight	1	1	1
Max Steps	$400K$	1M	1M
Max Epochs	$10K$	300	300
Learning Rate	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$2 \cdot 10^{-4}$
Batch Size	256	128	128
Warm-up Steps	$40K$	60K	60K

Table 2: Results on ZINC dataset. **Red**: best model, **Violet**: SOTA baseline model

Methods	#param	Test MAE
GIN [22]	$509,549$	0.526 ± 0.051
GAT [21]	$531,345$	0.398 ± 0.007
GCN [20]	$505,341$	0.367 ± 0.011
MPNN [2]	$480,805$	0.145 ± 0.007
PNA [27]	$387,155$	0.142 ± 0.010
GSN [28]	–	0.101 ± 0.010
CRAWL [24]	–	$\mathbf{0.085 \pm 0.004}$
GT [26]	$588,929$	0.226 ± 0.014
SAN [25]	$508,577$	0.139 ± 0.006
Graphormer [8]	$489,321$	0.122 ± 0.006
EGT [9]	–	0.108 ± 0.009
Ours	$419,729$	$\mathbf{0.085 \pm 0.003}$

Table 3: Results on PCQM4M-LSC dataset. **Red**: best model, **Violet**: SOTA baseline model.

Methods	#param	Valid MAE
GCN [20]	2.0M	0.1684
GIN [22]	3.8M	0.1536
GCN-VN [20]	4.9M	0.1510
GIN-VN [22]	6.7M	0.1396
DeeperGCN-VN [29]	25.5M	0.1398
MoleculeX [23]	34.1M	0.1278
GT [26]	0.6M	0.1400
Graphormer$_{SMALL}$[8]	12.5M	0.1264
Graphormer [8]	47.1M	0.1234
EGT [9]	47.4M	**0.1224**
Ours$_{SMALL}$	12.7M	0.1245
Ours	47.3M	**0.1223**

Results Table 2 shows the performance on ZINC. It can be observed that our model surpasses most models and approximates CRAWL with lower variance. Table 3 also summarizes performance on PCQM4M-LSC and our model performs better than Graphormer [8] with similar parameter quantity.

4.2 Ablation Study

We perform a series of ablation studies on the ZINC dataset to illustrate each component's effect in the proposed methods. The results are shown in Table 4.

Spatial Embedding We compare atom-based spatial embedding to the proposed motif-based spatial embedding. Although motif-based SPD introduces structural debilitating, it outperforms the vanilla atom-based SPD. This phenomenon shows that motif-based SPD is suitable for molecular graphs, and aromatic connectivity is vital for distance definition.

Multi-Scale Self-Attention To illustrate the importance of distance-guided multi-scale information, we compare distance-guided multi-scale self-attention with complete scales in all layers. The result shows that graph Transformer architecture with multi-scale self-attention yields a large margin of performance improvement.

Table 4: Ablation study results on ZINC dataset.

Spatial Embedding		Multi-Scale	Test MAE
atom-based	motif-based		
✓	-	-	0.122 ± 0.006
-	✓	-	0.098 ± 0.002
-	✓	✓	$\mathbf{0.085 \pm 0.003}$

4.3 Structural Debilitating Problem

We analyze the debilitating structural problem discussed above. We conduct ablation experiments with models with and without multi-scale self-attention on different diameter graphs. The results in Table 5, show that multi-scale self-attention significantly moderates the inconsistent performance on various diameter graphs. Especially for shorter distances, with relatively more vital interatomic forces, multi-scale self-attention achieves improvement more significantly.

5 Limitations

Although our methods achieve considerable progress, there are still some limitations. Firstly, we merely treat all kinds of rings as bi-connected components. The chemical properties of aromatic rings differ significantly from those of regular rings. Some sophisticated designs involving chemical bond

Table 5: Multi-scale self-attention on different diameter graphs.

Multi-scale	Diameter (MAE)								
	2	3	4	5	6	7	8	9	10
-	0.187	0.236	0.138	**0.176**	0.110	**0.064**	0.107	0.084	**0.057**
✓	**0.130**	**0.105**	**0.128**	0.186	**0.099**	0.066	**0.070**	**0.079**	0.061

properties as filtering signals could help us separate those two cases and further improve our results. Secondly, in our proposed methods, both edges represent the distance of a single unit, regardless of their chemical bond type. Specifically, multi-scale self-attention is motivated by interatomic forces. However, interatomic distance is related to bond length, bond angle, and dihedral angle [30], which we did not introduce into our method.

6 Conclusions

This paper proposes a chemistry-guided transformer for the molecular graph forecasting task, which achieves state-of-art performance. It can be seen as an example of how simple chemical intuitions can significantly help us in chemical property forecasting tasks. To add chemistry-inspired inductive bias into the graph Transformer, we propose a motif-based spatial embedding to represent aromaticity-related connectivity. Moreover, we introduce a distance-guided multi-scale self-attention to consider interatomic forces and eliminate the debilitating structural problem. Lastly, experiments show that the proposed method considerably improves on two large benchmark datasets compared with state-of-the-art algorithms.

References

[1] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021.

[2] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. In *International conference on machine learning (ICML)*, 2017.

[3] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. DeepGCNs: Can GCNs Go as Deep as CNNs? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9267–9276, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2016.

[5] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[6] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

[7] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys (CSUR)*, 2020.

[8] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[9] Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. Global self-attention as a replacement for graph convolution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 655–665, 2022.

[10] Alan J Rocke. It began with a daydream: the 150th anniversary of the kekulé benzene structure. *Angewandte Chemie International Edition*, 54(1):46–50, 2015.

[11] Anthony Stone. *The theory of intermolecular forces*. oUP oxford, 2013.

[12] Paul von Ragué Schleyer. Introduction: aromaticity. *Chemical Reviews*, 101(5):1115–1118, 2001.

[13] Janet E. Jones. On the determination of molecular fields. i. from the variation of the viscosity of a gas with temperature. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 106:441–462.

[14] Qipeng Guo, Xipeng Qiu, Pengfei Liu, Xiangyang Xue, and Zheng Zhang. Multi-Scale Self-Attention for Text Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7847–7854, 2020.

[15] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.

[16] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A large-scale challenge for machine learning on graphs. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[17] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[18] John Hopcroft and Robert Tarjan. Efficient algorithms for graph manipulation. *Communications of the ACM*, 16(6):372–378, 1973.

[19] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *arXiv preprint arXiv:2003.00982*, 2020.

[20] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[21] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[22] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*, 2019.

[23] Meng Liu, Cong Fu, Xuan Zhang, Limei Wang, Yaochen Xie, Hao Yuan, Youzhi Luo, Zhao Xu, Shenglong Xu, and Shuiwang Ji. Fast quantum property prediction via deeper 2d and 3d graph networks. In *NeurIPS 2021 AI for Science Workshop*, 2021.

[24] Jan Toenshoff, Martin Ritzert, Hinrikus Wolf, and Martin Grohe. Graph learning with 1d convolutions on random walks. *arXiv preprint arXiv:2102.08786*, 2021.

[25] Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:21618–21629, 2021.

[26] Vijay Prakash Dwivedi and Xavier Bresson. A Generalization of Transformer Networks to Graphs. In *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

[27] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal Neighbourhood Aggregation for Graph Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13260–13271, 2020.

[28] Giorgos Bouritsas, Fabrizio Frasca, Stefanos P Zafeiriou, and Michael Bronstein. Improving graph neural network expressivity via subgraph isomorphism counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[29] Guohao Li, Chenxin Xiong, Ali Thabet, and Bernard Ghanem. DeeperGCN: All you need to train deeper gcns. *arXiv preprint arXiv:2006.07739*, 2020.

[30] Josef Michl. Organic chemical systems, theory. In Robert A. Meyers, editor, *Encyclopedia of Physical Science and Technology (Third Edition)*, pages 435–457. Academic Press, New York, third edition edition, 2003.