Where to Mask: Structure-Guided Masking for Graph Masked Autoencoders

Chuang Liu[1], **Yuyao Wang**[1], **Yibing Zhan**[2], **Xueqi Ma**[3], **Dapeng Tao**[4], **Jia Wu**[5], **Wenbin Hu**[1]*
[1]School of Computer Science, Wuhan University, Wuhan, China
[2]JD Explore Academy, JD.com, China
[3]School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia
[4]School of Computer Science, Yunnan University, Kunming, China
[5]School of Computing, Macquarie University, Sydney, Australia
{chuangliu, wyy0224, hwb}@whu.edu.cn, zhanyibing@jd.com, xueqim@student.unimelb.edu.au, dptao@ynu.edu.cn, jia.wu@mq.edu.au

Abstract

Graph masked autoencoders (GMAE) have emerged as a significant advancement in self-supervised pre-training for graph-structured data. Previous GMAE models primarily utilize a straightforward random masking strategy for nodes or edges during training. However, this strategy fails to consider the varying significance of different nodes within the graph structure. In this paper, we investigate the potential of leveraging the graph's structural composition as a fundamental and unique prior in the masked pre-training process. To this end, we introduce a novel structure-guided masking strategy (*i.e.*, StructMAE), designed to refine the existing GMAE models. StructMAE involves two steps: **1)** Structure-based Scoring: Each node is evaluated and assigned a score reflecting its structural significance. Two distinct types of scoring manners are proposed: predefined and learnable scoring. **2)** Structure-guided Masking: With the obtained assessment scores, we develop an easy-to-hard masking strategy that gradually increases the structural awareness of the self-supervised reconstruction task. Specifically, the strategy begins with random masking and progresses to masking structure-informative nodes based on the assessment scores. This design gradually and effectively guides the model in learning graph structural information. Furthermore, extensive experiments consistently demonstrate that our StructMAE method outperforms existing state-of-the-art GMAE models in both unsupervised and transfer learning tasks. Codes are available at https://github.com/LiuChuang0059/StructMAE.

1 Introduction

In domains, such as academic, social, and biological networks, graph-structured data often lacks labels. This problem is especially acute in biochemical graphs due to the expense of wet-lab experiments. To address this challenge, many techniques have been developed to fully exploit the existing massive amounts of unlabeled data, aiming to enhance graph model training. Among these approaches, self-supervised graph pre-training (SSGP) methods are prominent due to their effectiveness, attracting significant interest in academic and industrial realms [Xia *et al.*, 2022b].

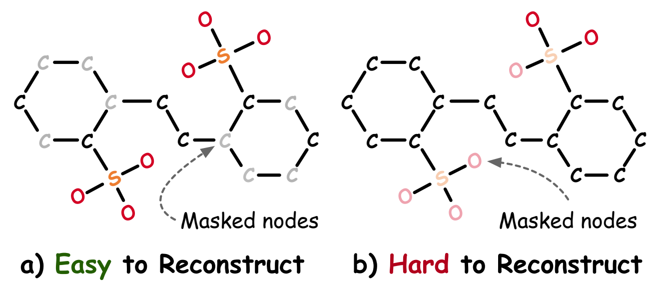

Figure 1: Two primary examples that underscore the potential sub-optimal nature of the *random masking* strategy in GMAE. **a)** Masked nodes are too simplistic to predict (*i.e.*, C), hindering the acquisition of valuable knowledge. **b)** Masking a large number of informative chemical nodes (*i.e.*, SO_3) makes the model fail to perceive the structural information in graphs.

Currently, SSGP methods are categorized into two primary streams: **1) Contrastive** methods, such as GraphCL [You *et al.*, 2020] and SimGRACE [Xia *et al.*, 2022a], which utilize contrastive learning principles to reveal the intrinsic structure and interconnections within graph data. **2) Generative** methods, including GraphMAE [Hou *et al.*, 2022] and MaskGAE [Li *et al.*, 2023]. These methods focus on learning node representations through a reconstruction objective. Moreover, generative methods have proved to be simpler and more effective than contrastive approaches that require carefully designed augmentation and sampling strategies. The efficacy of generative methods is further underscored by the enormous successes of models such as BERT and ChatGPT in Natural Language Processing (NLP) [Devlin *et al.*, 2019] and MAE in Computer Vision (CV) [He *et al.*, 2022]. These

*Corresponding Author

successes highlight the significant potential of generative approaches across various domains. Accordingly, this paper explores the capabilities of generative methods, specifically graph masked autoencoders (GMAE), in graph learning tasks and recognizes their potential as evidenced in other fields.

GMAE fundamentally involves randomly masking a proportion of input data (*i.e.*, nodes or edges) and leveraging the reconstruction of the removed contents to guide the representation learning. Despite GMAE's promising results, its random masking approach, which assigns equal probability to all nodes in a graph - a universally adopted strategy, presents a suboptimal strategy. Specifically, the masked nodes are sometimes overly simplistic to predict (*i.e.*, the atom C in Figure 1 (a)) with only neighborhood information. In such cases, the model's pre-training phase may not be sufficiently informative, thereby hindering the acquisition of valuable knowledge. However, if we mask a large number of key informative nodes (*i.e.*, SO_3 in Figure 1 (b)), the model may fail to perceive the graph's overall structural information. In summary, the indiscriminate nature of random masking, which fails to distinguish nodes of varying informational values, potentially leads to low data efficiency and compromises the quality of the learned graph representations. Therefore, this raises the question: *is there a more effective masking strategy than random sampling for enhancing GMAE's pre-training process?*

This paper answers the aforementioned open question by introducing StructMAE, which features a novel structure-guided masking strategy designed to enhance GMAE's pre-training process. The key insight of our method is to inject prior graph structure knowledge into the masking process to guide model learning. Specifically, StructMAE comprises two principal components: **Structure-based Scoring** and **Structure-guided Masking**. We recognize that the node reconstruction complexity is inherently linked to its structural significance within the graph. Therefore, we derive a scoring method to assess the node significance, distinguishing between informative and less-informative nodes based on structural considerations. In addition, two scoring method variants are proposed: predefined and learnable, which are discussed in detail in Section 4. With node significance scores established, we propose an easy-to-hard masking strategy that gradually increases the difficulty of the self-supervised reconstruction task. This approach commences with the masking of less-informative nodes, progressively shifting towards masking more informative nodes as the model's learning progresses. This strategic progression in masking difficulty is designed to enable the model to gradually and effectively assimilate the graph's structural information.

To evaluate the effectiveness of the StructMAE model, we conduct comprehensive experiments on a range of widely-used datasets, notably the large-scale Open Graph Benchmark (OGB) [Hu *et al.*, 2020a], covering two graph learning tasks: unsupervised and transfer learning. The experimental results consistently demonstrate that StructMAE's performance surpasses that of existing state-of-the-art models in contrastive and generative pre-training domains. This exceptional performance demonstrates our structure-guided masking approach's advantages over conventional random masking methods. The principal contributions are summarized as follows:

1. We introduce StructMAE, a novel node masking strategy tailored for GMAE. This strategy utilizes the structural information inherent in graphs to gradually and effectively direct the masking process, significantly enhancing the model's capability for representation learning.
2. We evaluate StructMAE through extensive experiments, comparing its performance with generative and contrastive baselines across two graph tasks on various real-world graph datasets, including the OGB dataset. The experimental results consistently validate StructMAE's effectiveness.

2 Related Work

Self-supervised Graph Pre-training. Inspired by the success of pre-trained language models like BERT [Devlin *et al.*, 2019], T5 [Raffel *et al.*, 2020], and ChatGPT [Brown *et al.*, 2020], numerous efforts have been directed towards SSGP. Based on model architectures and objective designs, SSGP is naturally divided into contrastive and generative methods. First, contrastive self-supervised learning has dominated graph representation learning in the past two years. Its success is largely due to elaborate data augmentation designs, negative sampling, and contrastive loss. For instance, DGI [Veličković *et al.*, 2019] and InfoGraph [Sun *et al.*, 2020], based on mutual information, leverage corruptions to construct negative pairs. Similarly, models such as SimGRACE [Xia *et al.*, 2022a] and GraphCL [You *et al.*, 2020] utilize in-batch negatives. Differing from previous methods that execute augmentation on graphs, COSTA [Zhang *et al.*, 2022b] implements augmentation within the embedding space, aiding in the mitigation of sampling bias. Moreover, POT [Yu *et al.*, 2024] advances graph contrastive learning (GCL) training through the employment of a node compactness metric, assessing adherence to the GCL principle. Second, generative self-supervised learning focuses on recovering missing parts of the input data. For example, GAE [Kipf and Welling, 2016] is a conventional method that reconstructs the adjacency matrix. Multiple GAE variants utilize graph reconstruction to pre-train Graph Neural Networks (GNNs), including ARVGA [Pan *et al.*, 2018] and SIGVAE [Hasanzadeh *et al.*, 2019]. Recently, a paradigm shift towards GMAE has shown promising results in various tasks [Liu *et al.*, 2023b]. A detailed introduction will be provided in the following section.

Graph Masked Autoencoders. GMAE mainly involves reconstructing the contents (*e.g.*, nodes and edges) that are randomly masked from the input using autoencoder architecture. A notable example of GMAE is GraphMAE [Hou *et al.*, 2022], which reconstructs randomly masked input node features with several innovative designs, including re-mask decoding and scaled cosine error. In addition, MaskGAE [Li *et al.*, 2023], S2GAE [Tan *et al.*, 2023], and GiGaMAE [Shi *et al.*, 2023] jointly reconstruct the masked edges and node degrees. SimSGT [Liu *et al.*, 2023c], and GCMAE [Wang *et al.*, 2024] combine contrastive learning with GMAE, whereas RARE [Tu *et al.*, 2024] employs self-distillation to enhance GMAE's performance. Unlike the above mentioned GMAE methods, which primarily use message-passing GNNs as backbone models, GMAE-GT [Zhang *et al.*, 2022a] utilizes a graph transformer [Ying *et al.*, 2021;

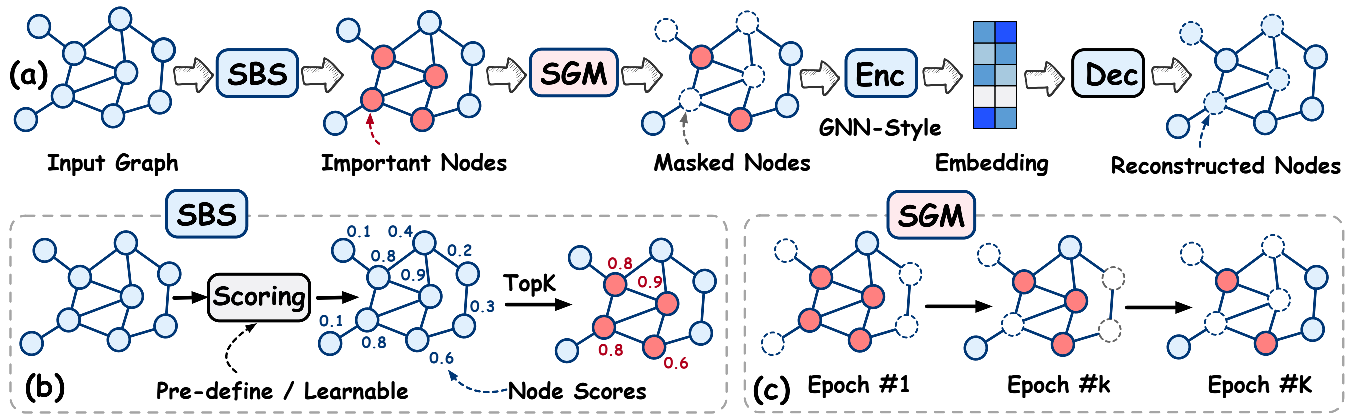

Figure 2: Overview of the proposed model. **(a)** The overall architecture of the proposed StructMAE. **(b)** SBS: It evaluates node importance based on the graph's structural information. This evaluation can be conducted using either a predefined or learnable approach. **(c)** SGM: It progressively increases the masking probability of important nodes as the training epochs advance.

Liu *et al.*, 2023a] as its encoder backbone. There are also several GMAE applications, including heterogeneous graph representation learning [Tian *et al.*, 2023], protein surface prediction [Yuan *et al.*, 2023], and action recognition [Yan *et al.*, 2023]. However, all the aforementioned methods employ a random masking method when masking graph contents and thus overlook the importance of GMAE mask strategies, potentially inhibiting the model's capabilities. The work most closely related to this study is MoAma [Inae *et al.*, 2023], which similarly focuses on designing superior mask strategies. However, this approach uses motifs which rely on domain knowledge and manual motif pre-definition, limiting its generalizability across various domains.

3 Preliminaries

Notations. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be represented by an adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ and a node feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, where $\mathcal{V}$ denotes the node sets, $\mathcal{E}$ denotes the edge sets, n is the number of nodes, d is the dimension of the node features, and $\mathbf{A}[i,j] = 1$ if there exists an edge between nodes v_i and v_j, otherwise, $\mathbf{A}[i,j] = 0$.

Graph Masked Autoencoders. GMAE operates as a self-supervised pre-training framework, which focuses on recovering masked node features or edges based on the representations of unmasked nodes. To illustrate this method, we focus on the reconstruction of node features as an example. GMAE comprises two essential components: an encoder ($f_E(\cdot)$) and a decoder ($f_D(\cdot)$). The encoder maps each unmasked node $v \in \mathcal{V}_{\text{unmask}}$ to a d-dimensional vector $\mathbf{h}_v \in \mathbb{R}^d$, with $\mathcal{V}_{\text{unmask}}$ representing the set of unmasked nodes, while the decoder reconstructs the masked node features from these vectors. The entire process can be formally represented as:

$$\mathbf{H}_{\text{unmask}} = f_E(\mathbf{A}, \mathbf{X}_{\text{unmask}}); \quad \mathbf{X}' = f_D(\mathbf{A}, \mathbf{H}_{\text{unmask}}), \tag{1}$$

where $\mathbf{X}_{\text{unmask}}$ and $\mathbf{H}_{\text{unmask}}$ denote the features and embeddings of unmasked nodes, respectively, and $\mathbf{X}'$ represents the reconstructed features of all the nodes. Then, GMAE optimizes the model by minimizing the discrepancy between the reconstructed representation of masked nodes, $\mathbf{X}'_{\text{mask}} \subset \mathbf{X}'$, and their original features, $\mathbf{X}_{\text{mask}} \subset \mathbf{X}$.

4 Methodology

This section presents the model architecture of StructMAE in detail. First, we provide a detailed introduction to the masking module in GMAE (§**4.1**). Subsequently, a comprehensive exploration (§**4.2**) of the masking strategy is provided. Then, the details of the proposed StrucMAE are elucidated (§**4.3**). Finally, the overall StructMAE architecture, encompassing training and inference details, is expounded (§**4.4**).

4.1 Introducing the GMAE Masking Module

Prior to processing graph data within the GMAE encoder, a subset of nodes $\mathcal{V}_{\text{mask}} \subset \mathcal{V}$ is sampled for masking. According to the methodology described in [Hou *et al.*, 2022], the features of these sampled nodes are replaced with a learnable vector $\boldsymbol{x}_{[\text{M}]} \in \mathbb{R}^d$. Accordingly, for a node v_i within the masked node subset $\mathcal{V}_{\text{mask}}$, its feature $\widetilde{\mathbf{x}}_i$ in the altered feature matrix $\widetilde{\mathbf{X}}$ is defined as follows:

$$\widetilde{\mathbf{x}}_i = \begin{cases} \mathbf{x}_{[\text{M}]} & v_i \in \mathcal{V}_{\text{mask}} \\ \mathbf{x}_i & v_i \notin \mathcal{V}_{\text{mask}} \end{cases}. \tag{2}$$

Regarding the approach for selecting nodes to mask, most existing works [Hou *et al.*, 2022; Li *et al.*, 2023] utilize a random masking strategy. This method assigns an equal masking probability of masking to each node within a graph. A more detailed exploration of the random masking strategy will be presented in the following section (§**4.2**).

4.2 Reconsidering the Random Masking Strategy

In GMAE context, the masking strategy is crucial, as it significantly influences the type of information that the model learns. In previous studies [Hou *et al.*, 2022], nodes within a graph are masked randomly, each with an equal probability. However, this strategy overlooks the varying structural information of different nodes, which has been shown to be crucial in graph learning tasks. Therefore, we conduct a preliminary

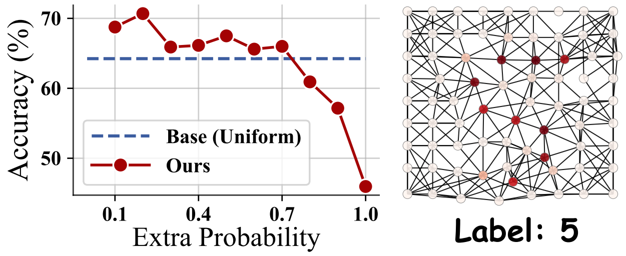

Figure 3: Effects of raising the masking probability on nodes with structural information (dark red nodes in the right part). The blue dashed line illustrates the results under the random masking strategy.

experiment to explore whether the incorporation of structure priors could augment pre-training efficacy.

In the experiment, models are pre-trained on the MNIST dataset and evaluated in unsupervised settings. Nodes forming numerical values are identified as those possessing rich structural information. Subsequently, the masking probability for these structurally informative nodes is manually increased. The results, presented on the left side of Figure 3, reveal the impact of this modified masking strategy on unsupervised accuracy. The results indicate that: **1)** A marginal increase in the masking probability for nodes with rich structural information enhances the model's pre-training learning. Furthermore, up to a probability threshold of 0.2, a corresponding gradual increase in the model's accuracy is observed. **2)** Conversely, excessively raising the masking probability of these structurally significant nodes detrimentally affects the model's training. These findings corroborate our initial discussion and highlight the importance of proposing an effective method to integrate structural information into the masking process.

4.3 The Proposed Masking Strategy

Inspired by the preceding discussion, we introduce an innovative structure-guided masking approach for GMAE, named StructMAE (as depicted in Figure 2 (b)). StructMAE involves integrating the graph's structural knowledge into the masking process, thereby directing the model's learning trajectory more effectively. StructMAE is composed of two primary elements: **Structure-based Scoring (SBS)** and **Structure-guided Masking (SGM)**.

Structure-based Scoring

The SBS evaluates the significance of each node based on its structural role within the graph. This evaluation facilitates the identification of nodes that are pivotal for the model to learn, allowing for a more targeted masking approach. To determine the importance of nodes, we introduce two distinct methodologies: the pre-defined and learnable methods.

Pre-defined Structure-based Scoring. The predefined method involves using a set of predetermined criteria, based on the known structural information, to evaluate node importance. Specifically, the computation of importance score $\mathbf{S} \in \mathbb{R}^n$ is achieved using the PageRank algorithm [Page *et al.*, 1999], a well-established technique for evaluating node significance based on graph structures. Thus, the importance score of node v_i is defined as:

$$s_i = \frac{1-e}{n} + e \sum_{j \in \mathcal{N}_i} \frac{s_j}{L_j}, \tag{3}$$

where e denotes the damping factor, L_i represents the degree of node v_i, and $\mathcal{N}_i$ denotes the set of neighboring nodes of node v_i. In addition to PageRank, other prevalent metrics for assessing node importance comprise degree, closeness centrality, and betweenness centrality. Each of these methods is based on different underlying principles, offering diverse perspectives on a node's role and influence within a graph. Although these methods constitute more intricate ways of evaluating node importance, our empirical findings suggest that PageRank serves as a straightforward and effective measure. A detailed discussion and comparison of these methods is presented in the following section.

Learnable Structure-based Scoring. In contrast to the predefined method, the learnable approach dynamically assesses node significance based on the evolving state of the graph during the learning process. Specifically, it integrates the formulation of the assessment metric with masked graph modeling, enabling end-to-end learning of this metric. To accomplish this, we employ a lightweight scoring network, denoted as $f_S(\cdot)$, which assesses the importance of each node v_i. The scoring network's design is similar to a GNN-style layer, and it effectively captures graph structural information. Thus, the importance score $s_i \in \mathbf{S}$ for node v_i is calculated as follows:

$$s_i = \text{Sigmoid}\left(f_S\left(\mathbf{x}_i, \mathbf{A}\right)\right), \quad i = 1, \ldots, n. \tag{4}$$

A higher score s_i indicates greater importance of the corresponding node v_i. Following the scoring, node features are sorted in descending order based on these scores. The ordered node features and their respective scores are represented as $\{\mathbf{x}'_i\}$ and $\{s'_i\}$, respectively, where $i = 1, \ldots, n$. To facilitate learning the scoring network $f_S(\cdot)$, the predicted scores are multiplied by the node features to serve as modulating factors. This operation is formally expressed as:

$$\hat{\mathbf{X}} = \left\{\mathbf{x}''_i \mid \boldsymbol{x}''_i = \boldsymbol{x}'_i * s'_i\right\}, \quad i = 1, \ldots, n, \tag{5}$$

where $\hat{\mathbf{X}}$ denotes the set of node features after scoring. This mechanism ensures that the scoring network is continually updated and refined throughout the model's training process.

In this section, we explore the two distinct SBS methods proposed for StructMAE. First, the pre-defined SBS method offers simplicity and is particularly effective when the structural characteristics are well-understood and can be explicitly defined beforehand. Conversely, the learnable SBS is more adaptable and can cater to complex and variable graph structures, making it suitable for scenarios where the node significance cannot be easily predetermined.

Structure-guided Masking

The SGM component utilizes the scores generated by the SBS to guide its masking decisions. It selectively and progressively masks nodes in an easy-to-hard manner, thereby enhancing the model's capacity to effectively learn and represent graph structures. Specifically, in the initial training stage, a subset of easy nodes with lower scores is masked, making it easier for the model to predict them using basic neighboring information. As training progresses, the masking strategy evolves to encompass more challenging nodes. This enables the model to capture intricate structural information and, consequently, enhances its learning capabilities [Liu *et al.*, 2017; Liu *et al.*, 2018].

This masking strategy relies on the importance scoring matrix $\mathbf{S}$. It is hypothesized that nodes with higher $\mathbf{S}$ scores are more informative and significant. Consequently, we gradually increase the masking probabilities for these high-scoring nodes during masked graph modeling. To implement this, we rank the nodes based on their scores $\mathbf{S} = \{s_1, s_2, \cdots, s_n\}$ and identify the top K indices that constitute the set $\mathcal{Y}$. The masking probability for each node is then determined as:

$$\gamma_i = \epsilon + \begin{cases} \beta & v_i \in \mathcal{Y} \\ 0 & v_i \notin \mathcal{Y} \end{cases}, \tag{6}$$

where ϵ denotes random noise drawn from a uniform distribution ($\epsilon \sim U(0, 1)$) and β represents the increased probability assigned to nodes with higher scores. In this instance, nodes in the set $\mathcal{Y}$ are considered more informative, and consequently, the model is anticipated to prioritize these nodes. The number of masked nodes, denoted by K, is dynamically adjusted throughout the training process. Initially set at zero, K progressively increases with epoch according to the following formula:

$$K(t) = p \cdot n \cdot \sqrt{t/T}, \tag{7}$$

where $K(t)$ represents the K value at epoch t, n denotes the total number of nodes in the graph, p is the predefined mask ratio, and T represents the total number of training epochs. This approach enables the model to progressively concentrate on more challenging nodes, thereby enhancing its acquisition of complex structural information.

4.4 Overall StructMAE Architecture

Training Process. The StructMAE training process begins with an input graph from which a specified proportion of nodes is chosen based on our selective masking strategy. The selected nodes are subsequently masked using a mask-token. The graph, now with partially masked features, is subsequently fed into the encoder, which generates encoded representations of the nodes. After that, the decoder module is responsible for predicting and reconstructing the features of masked nodes. For optimization, we adopt the scaled cosine error as utilized in GraphMAE [Hou *et al.*, 2022].

Inference and Downstream Tasks. StructMAE is designed to cater to two distinct downstream applications: unsupervised and transfer learning. In unsupervised learning, the encoder processes the input graph without masking during the inference stage. The node embeddings generated by the encoder are then utilized for graph classification tasks with linear classifiers or support vector machines. In the transfer learning context, the pre-trained models are fine-tuned on different datasets. This fine-tuning enables the model to adjust to new data domains, leveraging the foundational knowledge gained during its initial training on the source dataset. Each of these downstream tasks emphasizes the versatility and applicability of StructMAE in diverse graph learning scenarios.

5 Experiment

5.1 Unsupervised Representation Learning

Objective. To assess the efficacy of the pre-trained model in its feature extraction capability, we subject it to a series of unsupervised tasks. Achieving success in these tasks will highlight the model's proficiency in learning high-quality and informative representations, which are crucial for various downstream graph analytics tasks.

Settings. We employ seven real-world datasets, including MUTAG, IMDB-B, IMDB-M, PROTEINS, COLLAB, REDDIT-B, and NCI1, involving diverse domains and sizes. To demonstrate the effectiveness of our proposed method, we compare StructMAE with the following 10 baseline models: *a) Two supervised models:* GIN [2019] and DiffPool [2018]; *b) Six contrastive models:* Infograph [2020], GraphCL [2020], JOAO [2021], GCC [2020], InfoGCL [2021a], and SimGRACE [Xia *et al.*, 2022a]; *c) Two generative models:* GraphMAE [Hou *et al.*, 2022] and S2GAE [Tan *et al.*, 2023]. We report the results from previous papers according to graph classification research norms. In the evaluation protocol, we initially generate graph embeddings using the encoder and readout function. Then, the encoded graph-level representations are fed into a downstream LIBSVM [Chang and Lin, 2011] classifier for label prediction, consistent with other baseline models. The performance is assessed by measuring the mean accuracy obtained from a 10-fold cross-validation, and this evaluation is repeated five times to ensure robustness.

Results. The results are detailed in Table 1, where StructMAE-P denotes StructMAE with predefined SBS, and StructMAE-L is with learnable SBS. Analyzing these results enables deriving several observations. **1)** State-of-the-art Performance: StructMAE-L outperforms existing self-supervised baselines on four out of seven datasets. Furthermore, it attains state-of-the-art performance considering the average rank across these datasets. Meanwhile, StructMAE-P maintains competitive performance against other self-supervised methods. These results emphatically demonstrates the efficacy of the proposed StructMAE approach. Please note that, StructMAE focuses solely on masking nodes, whereas S2GAE extends its masking strategy to include edges. The differing approach of S2GAE underscores a potential avenue for further development in StructMAE. **2)** Comparison with Supervised Methods: Remarkably, the StructMAE-L, though self-supervised, attain comparable or superior performance on certain datasets, including PROTEINS, IMDB-B, and COLLAB. This finding indicates that the representations learned by StructMAE are high-quality and informative, aligning with supervised learning benchmarks. **3)** Comparison with GraphMAE: In comparison to GraphMAE, which employs a random node masking strategy, StructMAE consistently outperforms it, providing further evidence to support our hypothesis that incorporating structural knowledge into the masking process can significantly enhance the model's learning capabilities. **4)** Predefined vs. Learnable SBS: A notable trend is that the StructMAE-L's performance generally surpasses that of StructMAE-P, particularly on complex datasets such as COLLAB and REDDIT-B. This trend indicates the greater

	PROTEINS	**NCI1**	**IMDB-B**	**IMDB-M**	**COLLAB**	**REDDIT-B**	**MUTAG**	**A.R.**
				Supervised Methods				
GIN [Xu *et al.*, 2019]	$76.2_{\pm 2.8}$	$82.7_{\pm 1.7}$	$75.1_{\pm 5.1}$	$52.3_{\pm 2.8}$	$80.2_{\pm 1.9}$	$92.4_{\pm 2.5}$	$89.4_{\pm 5.6}$	–
DiffPool [Ying *et al.*, 2018]	–	$92.1_{\pm 2.6}$	$72.6_{\pm 3.9}$	–	$78.9_{\pm 2.3}$	$92.1_{\pm 2.6}$	$75.1_{\pm 3.5}$	–
				Self-supervised Methods				
Infograph [Sun *et al.*, 2020]	$74.44_{\pm 0.31}$	$76.20_{\pm 1.06}$	$73.03_{\pm 0.87}$	$49.69_{\pm 0.53}$	$70.65_{\pm 1.13}$	$82.50_{\pm 1.42}$	$89.01_{\pm 1.13}$	6.86
GraphCL [You *et al.*, 2020]	$74.39_{\pm 0.45}$	$77.87_{\pm 0.41}$	$71.14_{\pm 0.44}$	$48.58_{\pm 0.67}$	$71.36_{\pm 1.15}$	$89.53_{\pm 0.84}$	$86.80_{\pm 1.34}$	7.43
JOAO [You *et al.*, 2021]	$74.55_{\pm 0.41}$	$78.07_{\pm 0.47}$	$70.21_{\pm 3.08}$	$49.20_{\pm 0.77}$	$69.50_{\pm 0.36}$	$85.29_{\pm 1.35}$	$87.35_{\pm 1.02}$	8.00
GCC [Qiu *et al.*, 2020]	–	–	72.0	49.4	78.9	**89.8**	–	5.25
InfoGCL [Xu *et al.*, 2021a]	–	$80.20_{\pm 0.60}$	$75.10_{\pm 0.90}$	$51.40_{\pm 0.80}$	$80.00_{\pm 1.30}$	–	$\mathbf{91.20_{\pm 1.30}}$	4.00
SimGRACE [Xia *et al.*, 2022a]	$75.35_{\pm 0.09}$	$79.12_{\pm 0.44}$	$71.30_{\pm 0.77}$	–	$71.72_{\pm 0.82}$	$89.51_{\pm 0.89}$	$89.01_{\pm 1.31}$	5.00
GraphMAE [Hou *et al.*, 2022]	$75.30_{\pm 0.39}$	$80.40_{\pm 0.30}$	$75.52_{\pm 0.66}$	$51.63_{\pm 0.52}$	$80.32_{\pm 0.46}$	$88.01_{\pm 0.19}$	$88.19_{\pm 1.26}$	4.43
S2GAE [Tan *et al.*, 2023]	$76.37_{\pm 0.43}$	$80.80_{\pm 0.24}$	$75.76_{\pm 0.62}$	$51.79_{\pm 0.36}$	$81.02_{\pm 0.53}$	$87.83_{\pm 0.27}$	$88.26_{\pm 0.76}$	3.14
StructMAE-P (ours)	$75.97_{\pm 0.38}$	$\mathbf{81.91_{\pm 0.31}}$	$75.72_{\pm 0.36}$	$51.25_{\pm 0.64}$	$80.53_{\pm 0.22}$	$88.25_{\pm 0.40}$	$87.91_{\pm 0.39}$	3.71
StructMAE-L (ours)	$\mathbf{76.62_{\pm 0.84}}$	$81.25_{\pm 1.37}$	$\mathbf{75.84_{\pm 0.46}}$	$\mathbf{52.05_{\pm 0.73}}$	$\mathbf{81.46_{\pm 0.27}}$	$89.03_{\pm 0.40}$	$88.43_{\pm 0.54}$	**1.86**

Table 1: Experimental results for **unsupervised representation learning** in graph classification. The results for baseline methods are sourced from prior studies. **Bold** or underline indicates the best or second-best result, respectively, among self-supervised methods. **A.R.** denotes the average rank of self-supervised methods.

	BBBP	**Tox21**	**ToxCast**	**SIDER**	**ClinTox**	**MUV**	**HIV**	**BACE**	**Avg.**
No-pretrain	$65.5_{\pm 1.8}$	$74.3_{\pm 0.5}$	$63.3_{\pm 1.5}$	$57.2_{\pm 0.7}$	$58.2_{\pm 2.8}$	$71.7_{\pm 2.3}$	$75.4_{\pm 1.5}$	$70.0_{\pm 2.5}$	67.0
ContextPred [Hu* *et al.*, 2020b]	$64.3_{\pm 2.8}$	$75.7_{\pm 0.7}$	$63.9_{\pm 0.6}$	$60.9_{\pm 0.6}$	$65.9_{\pm 3.8}$	$75.8_{\pm 1.7}$	$77.3_{\pm 1.0}$	$79.6_{\pm 1.2}$	70.4
AttrMasking [Hu* *et al.*, 2020b]	$64.3_{\pm 2.8}$	$76.7_{\pm 0.4}$	$64.2_{\pm \mathbf{0.5}}$	$61.0_{\pm 0.7}$	$71.8_{\pm 4.1}$	$74.7_{\pm 1.4}$	$77.2_{\pm 1.1}$	$79.3_{\pm 1.6}$	71.1
Infomax [Hu* *et al.*, 2020b]	$68.8_{\pm 0.8}$	$75.3_{\pm 0.5}$	$62.7_{\pm 0.4}$	$58.4_{\pm 0.8}$	$69.9_{\pm 3.0}$	$75.3_{\pm 2.5}$	$76.0_{\pm 0.7}$	$75.9_{\pm 1.6}$	70.3
GraphCL [You *et al.*, 2020]	$69.7_{\pm 0.7}$	$73.9_{\pm 0.7}$	$62.4_{\pm 0.6}$	$60.5_{\pm 0.9}$	$76.0_{\pm 2.7}$	$69.8_{\pm 2.7}$	$\mathbf{78.5_{\pm 1.2}}$	$75.4_{\pm 1.4}$	70.8
JOAO [You *et al.*, 2021]	$70.2_{\pm 1.0}$	$75.0_{\pm 0.3}$	$62.9_{\pm 0.5}$	$60.0_{\pm 0.8}$	$81.3_{\pm 2.5}$	$71.7_{\pm 1.4}$	$76.7_{\pm 1.2}$	$77.3_{\pm 0.5}$	71.9
GraphLoG [Xu *et al.*, 2021b]	$72.5_{\pm 0.8}$	$75.7_{\pm 0.5}$	$63.5_{\pm 0.7}$	$61.2_{\pm 1.1}$	$76.7_{\pm 3.3}$	$76.0_{\pm 1.1}$	$77.8_{\pm 0.8}$	$83.5_{\pm 1.2}$	73.4
RGCL [Li *et al.*, 2022]	$71.2_{\pm 0.9}$	$75.3_{\pm 0.5}$	$63.1_{\pm 0.3}$	$61.2_{\pm 0.6}$	$85.0_{\pm 0.8}$	$73.1_{\pm 1.2}$	$77.3_{\pm 0.8}$	$75.7_{\pm 1.3}$	72.7
GraphMAE [Hou *et al.*, 2022]	$72.0_{\pm 0.6}$	$75.5_{\pm 0.6}$	$64.1_{\pm 0.3}$	$60.3_{\pm 1.1}$	$82.3_{\pm 1.2}$	$76.3_{\pm 2.4}$	$77.2_{\pm 1.0}$	$83.1_{\pm 0.9}$	73.8
GraphMAE2 [Hou *et al.*, 2023]	$71.6_{\pm 1.6}$	$75.9_{\pm 0.8}$	$\mathbf{65.6_{\pm 0.7}}$	$59.6_{\pm 0.6}$	$78.8_{\pm 3.0}$	$78.5_{\pm 1.1}$	$76.1_{\pm 2.2}$	$81.0_{\pm 1.4}$	73.4
Mole-BERT [Xia *et al.*, 2023]	$71.9_{\pm 1.6}$	$\mathbf{76.8_{\pm 0.5}}$	$64.3_{\pm 0.2}$	$\mathbf{62.8_{\pm 1.1}}$	$78.9_{\pm 3.0}$	$\mathbf{78.6_{\pm 1.8}}$	$78.2_{\pm 0.8}$	$80.8_{\pm 1.4}$	74.0
StructMAE-P (ours)	$\mathbf{72.6_{\pm 0.9}}$	$75.8_{\pm 0.4}$	$64.5_{\pm 0.5}$	$62.0_{\pm 0.4}$	$86.0_{\pm 1.6}$	$77.7_{\pm 1.1}$	$77.4_{\pm 1.0}$	$\mathbf{84.3_{\pm 0.6}}$	75.0
StructMAE-L (ours)	$72.5_{\pm 0.9}$	$75.3_{\pm 0.4}$	$64.0_{\pm 0.4}$	$61.3_{\pm 0.5}$	$\mathbf{87.9_{\pm 2.1}}$	$78.0_{\pm 1.1}$	$78.3_{\pm 0.8}$	$83.2_{\pm 0.9}$	**75.1**

Table 2: Experimental results for **transfer learning** on molecular property prediction. The model is initially pre-trained on the ZINC15 dataset and subsequently fine-tuned on the above datasets. The reported metrics are ROC-AUC scores. The results for baseline methods are derived from prior studies. **Bold** or underline indicates the best or second-best result, respectively. **Avg.** denotes the average performance.

adaptability and effectiveness of the learnable SBS in handling complex and variable graph structures, as opposed to the predefined method. These results collectively validate the core principles behind StructMAE and its components, highlighting its potential as a powerful tool for unsupervised representation learning in graph-structured data.

5.2 Transfer Learning

Objective. The primary goal of the transfer learning task is to evaluate the transferability of the pre-training scheme utilized in StructMAE. This involves pre-training the model on a specific dataset and fine-tuning it with different datasets.

Metric	IM-M	COL
PageRank	**51.25**	**80.53**
Degree	50.87	80.28
Close.	51.04	80.44
Between.	50.97	80.17

(a) **Pre-defined metric.** Close. and Between. denote closeness centrality and betweenness centrality, respectively.

Function	IM-B	RE-B
MLP	75.32	88.70
GNN	75.52	88.83
M&G	**75.84**	**89.03**

(b) **Scoring function.** M&G mixes the scores generated by MLP and GNN.

Strategy	IM-B	RE-B
Top	75.62	87.81
Middle	75.08	75.03
Bottom	75.28	87.82
E-to-H	**75.72**	**88.35**

(c) **Masking strategy.** E-to-H is short for easy-to-hard.

Dataset	w/o	w
IM-B	74.92	**75.72**
IM-M	50.36	**51.25**
COL	79.27	**80.53**
RE-B	81.71	**88.25**

(d) **Randomness.** Performance of StructMAE with/without random masking.

Table 3: **Ablation Study**. The best results are in **bold**. Default settings are marked in gray. IM-B, IM-M, RE-B and COL correspond to IMDB-B, IMDB-M, REDDIT-B and COLLAB, respectively.

Settings. During the initial pre-training phase, StructMAE is trained on a dataset comprising two million unlabeled molecules obtained from the ZINC15 [Sterling and Irwin, 2015] dataset. Subsequently, the model is fine-tuned on eight classification benchmark datasets featured in the MoleculeNet dataset [Wu *et al.*, 2018]. In our evaluation, we adopt a scaffold-split approach for splitting the datasets, as outlined in [Hou *et al.*, 2022]. To demonstrate the effectiveness of our proposed method, we compare StructMAE with the following 10 baseline models: *a) Three unsupervised models:* Infomax, AttrMasking and ContextPred [Hu* *et al.*, 2020b]; *b) Four contrastive models:* GraphCL [2020], JOAO [2021], GraphLOG [Xu *et al.*, 2021b], and RGCL [Li *et al.*, 2022]; *c) Three generative models:* GraphMAE [Hou *et al.*, 2022], GraphMAE2 [Hou *et al.*, 2023], and Mole-BERT [Xia *et al.*, 2023]. We report the results of baseline models from previous papers according to research norms. Experiments are conducted 10 times, and the mean and standard deviation of the ROC-AUC scores are reported. According to the default settings used in prior research [Hou *et al.*, 2022], a 5-layer GIN model [Xu *et al.*, 2019] is employed as the encoder and a single-layer GIN as the decoder in StructMAE.

Results. The detailed results in Table 2 offer insightful observations into StructMAE's performance within the transfer learning context. **1)** State-of-the-Art Performance Across Datasets: StructMAE-P and StructMAE-L demonstrate state-of-the-art performance across an ensemble of eight datasets. Specifically, StructMAE-P and StructMAE-L achieved 1.4% and 1.5% improvements in average performance metrics, respectively. Additionally, each method individually achieves top-tier performance on several datasets. These results validate the StructMAE's ability to effectively generalize learned representations across diverse datasets. **2)** Superiority of StructMAE-L: The results reveal the superior performance of StructMAE-L over StructMAE-P on the average metrics. This trend further highlights the enhanced adaptability and effectiveness of the learnable SBS method, similar to observations made previously in unsupervised learning.

5.3 Ablation Study

A detailed study is conducted to evaluate the impact of different components within StructMAE. It is important to note that, except for the components under analysis, all other aspects of the model remain consistent with the comprehensive StructMAE. The findings, as outlined below, provide valuable insights into the significance of each component: **1)** Efficacy of Pre-defined Metrics: As shown in Table 3a, the PageRank metric consistently demonstrates superior performance compared to other pre-defined metrics. However, other metrics also demonstrate commendable performances, suggesting their potential applications in specific scenarios. **2)** Scoring Function Comparison: As displayed in Table 3b, we observe that the performance of GNN scoring function outperforms the Multilayer Perceptron (MLP). This outcome emphasizes the importance of incorporating structural information into the scoring process. Furthermore, the combined use of GNN and MLP consistently yields superior performance compared to using either one independently. **3)** Masking Strategy Efficiency: The results are detailed in Table 3c, where Top, Middle, and Bottom denote the masking of the top, middle, and bottom $p * n$ nodes, respectively, based on the masking probability. Analysis of the results indicates that the easy-to-hard masking strategy consistently surpasses the performance of other methods. This finding supports our perspective on the effectiveness of gradually increasing the learning challenge, validating the strategic design of our training approach. **4)** Role of Random Noise: Notably, a significant drop in performance is observed in the absence of random noise, as indicated in Table 3d. This indicates that the inclusion of randomness in the model enables access to a broader range of information and simultaneously strengthens the learning process robustness.

6 Conclusion

This study proposes the StructMAE model, a novel structure-guided masking strategy that incorporates prior structural knowledge into the masking process, thereby enhancing the pre-training model's learning efficiency. The StructMAE framework consists of two pivotal steps: SBS and SGM. Extensive experiments, encompassing two distinct graph learning tasks, demonstrate that StructMAE significantly outperforms existing self-supervised pre-training methods. These results highlight our approach's effectiveness in leveraging structural information for improved model performance. Despite its competitive performance, StructMAE still has room for improvement. For instance, 1) devising more effective scoring methods to fully exploit the structural information, 2) extending the structure-guided masking strategy to encompass edge masking, and 3) expanding structure-guided masking to a broader spectrum of tasks (*e.g.*, node classification).

Acknowledgments

This work was partially supported by the Major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" key project (No. 2021ZD0111700). The work of Wenbin Hu was supported by the National Key Research and Development Program of China (2023YFC2705700). This work was supported in part by the Natural Science Foundation of China (No. 82174230), Artificial Intelligence Innovation Project of Wuhan Science and Technology Bureau (No. 2022010702040070), Natural Science Foundation of Shenzhen City (No. JCYJ20230807090211021).

Contribution Statement

Chuang Liu and Yuyao Wang contributed equally to this work.

References

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *TIST*, 2011.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.

[Hasanzadeh *et al.*, 2019] Arman Hasanzadeh, Ehsan Hajiramezanali, Krishna Narayanan, Nick Duffield, Mingyuan Zhou, and Xiaoning Qian. Semi-implicit graph variational auto-encoders. In *NeurIPS*, 2019.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[Hou *et al.*, 2022] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *SIGKDD*, 2022.

[Hou *et al.*, 2023] Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. Graphmae2: A decoding-enhanced masked self-supervised graph learner. In *WWW*, 2023.

[Hu *et al.*, 2020a] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv:2005.00687*, 2020.

[Hu* *et al.*, 2020b] Weihua Hu*, Bowen Liu*, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *ICLR*, 2020.

[Inae *et al.*, 2023] Eric Inae, Gang Liu, and Meng Jiang. Motif-aware attribute masking for molecular graph pre-training. *arXiv:2309.04589*, 2023.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv:1611.07308*, 2016.

[Li *et al.*, 2022] Sihang Li, Xiang Wang, An Zhang, Yingxin Wu, Xiangnan He, and Tat-Seng Chua. Let invariant rationale discovery inspire graph contrastive learning. In *ICLR*, 2022.

[Li *et al.*, 2023] Jintang Li, Ruofan Wu, Wangbin Sun, Liang Chen, Sheng Tian, Liang Zhu, Changhua Meng, Zibin Zheng, and Weiqiang Wang. What's behind the mask: Understanding masked graph modeling for graph autoencoders. In *SIGKDD*, 2023.

[Liu *et al.*, 2017] Weiwei Liu, Ivor W Tsang, Müller Klaus-Robert, et al. An easy-to-hard learning paradigm for multiple classes and multiple labels. *JMLR*, 2017.

[Liu *et al.*, 2018] Weiwei Liu, Donna Xu, Ivor W Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE TPAMI*, 2018.

[Liu *et al.*, 2023a] Chuang Liu, Yibing Zhan, Xueqi Ma, Liang Ding, Dapeng Tao, Jia Wu, and Wenbin Hu. Gapformer: Graph transformer with graph pooling for node classification. In *IJCAI-23*, 2023.

[Liu *et al.*, 2023b] Chuang Liu, Yibing Zhan, Xueqi Ma, Dapeng Tao, Bo Du, and Wenbin Hu. Masked graph autoencoder constrained graph pooling. In *ECML PKDD*, 2023.

[Liu *et al.*, 2023c] Zhiyuan Liu, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Rethinking tokenizer and decoder in masked graph modeling for molecules. *NeurIPS*, 2023.

[Page *et al.*, 1999] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[Pan *et al.*, 2018] Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. Adversarially regularized graph autoencoder for graph embedding. In *IJCAI*, 2018.

[Qiu *et al.*, 2020] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *SIGKDD*, 2020.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[Shi *et al.*, 2023] Yucheng Shi, Yushun Dong, Qiaoyu Tan, Jundong Li, and Ninghao Liu. Gigamae: Generalizable graph masked autoencoder via collaborative latent space reconstruction. In *CIKM*, 2023.

[Sterling and Irwin, 2015] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 2015.

[Sun *et al.*, 2020] Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2020.

[Tan *et al.*, 2023] Qiaoyu Tan, Ninghao Liu, Xiao Huang, Soo-Hyun Choi, Li Li, Rui Chen, and Xia Hu. S2gae: Self-supervised graph autoencoders are generalizable learners with graph masking. In *WSDM*, 2023.

[Tian *et al.*, 2023] Yijun Tian, Kaiwen Dong, Chunhui Zhang, Chuxu Zhang, and Nitesh V Chawla. Heterogeneous graph masked autoencoders. In *AAAI*, 2023.

[Tu *et al.*, 2024] Wenxuan Tu, Qing Liao, Sihang Zhou, Xin Peng, Chuan Ma, Zhe Liu, Xinwang Liu, and Zhiping Cai. Rare: Robust masked graph autoencoder. *IEEE TKDE*, 2024.

[Veličković *et al.*, 2019] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *ICLR*, 2019.

[Wang *et al.*, 2024] Yuxiang Wang, Xiao Yan, Chuang Hu, Fangcheng Fu, Wentao Zhang, Hao Wang, Shuo Shang, and Jiawei Jiang. Generative and contrastive paradigms are complementary for graph self-supervised learning. In *ICDE*, 2024.

[Wu *et al.*, 2018] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 2018.

[Xia *et al.*, 2022a] Jun Xia, Lirong Wu, Jintao Chen, Bozhen Hu, and Stan Z. Li. Simgrace: A simple framework for graph contrastive learning without data augmentation. In *WWW*, 2022.

[Xia *et al.*, 2022b] Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z Li. A survey of pretraining on graphs: Taxonomy, methods, and applications. *arXiv:2202.07893*, 2022.

[Xia *et al.*, 2023] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *ICLR*, 2023.

[Xu *et al.*, 2019] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

[Xu *et al.*, 2021a] Dongkuan Xu, Wei Cheng, Dongsheng Luo, Haifeng Chen, and Xiang Zhang. InfoGCL: Information-aware graph contrastive learning. In *NeurIPS*, 2021.

[Xu *et al.*, 2021b] Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised graph-level representation learning with local and global structure. In *ICML*, 2021.

[Yan *et al.*, 2023] Hong Yan, Yang Liu, Yushen Wei, Zhen Li, Guanbin Li, and Liang Lin. Skeletonmae: graph-based masked autoencoder for skeleton sequence pre-training. In *ICCV*, 2023.

[Ying *et al.*, 2018] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, 2018.

[Ying *et al.*, 2021] 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 *NeurIPS*, 2021.

[You *et al.*, 2020] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

[You *et al.*, 2021] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *ICML*, 2021.

[Yu *et al.*, 2024] Yue Yu, Xiao Wang, Mengmei Zhang, Nian Liu, and Chuan Shi. Provable training for graph contrastive learning. *NeurIPS*, 36, 2024.

[Yuan *et al.*, 2023] Mingzhi Yuan, Ao Shen, Kexue Fu, Jiaming Guan, Yingfan Ma, Qin Qiao1, and Manning Wang. ProteinMAE: Masked Autoencoder for Protein Surface Self-supervised Learning. *Bioinformatics*, 2023.

[Zhang *et al.*, 2022a] Sixiao Zhang, Hongxu Chen, Haoran Yang, Xiangguo Sun, Philip S Yu, and Guandong Xu. Graph masked autoencoders with transformers. *arXiv:2202.08391*, 2022.

[Zhang *et al.*, 2022b] Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, and Irwin King. Costa: covariance-preserving feature augmentation for graph contrastive learning. In *SIGKDD*, pages 2524–2534, 2022.