ALTeGraD Data Challenge
Molecular Graph Captioning

Tristan Lecourtois
ENS Paris-Saclay
tristan.lecourtois@ens-paris-saclay.fr

Amine Ould
ENS Paris-Saclay
amine.ould_hocine@ens-paris-saclay.fr

Gloire Linvani
ENS Paris-Saclay
gloire.linvani@ens-paris-saclay.fr

1 Introduction

Multimodality in machine learning involves the interaction of information from different types of data, requiring models that can understand and relate these varied forms of input. In this project, we focus on the task of *Molecular Graph Captioning*, which aims to generate coherent and informative textual descriptions from molecular structures represented as graphs. The core challenge arises from the mismatch between the symbolic, relational nature of molecular graphs and the sequential, semantic structure of natural language.

Our objective is to design a model capable of translating atomic and bond-level information into human-readable captions that accurately summarize the key structural and functional properties of molecules. To address this challenge, we explored three fundamentally different paradigms:

(1) **Retrieval-Based Approach:** A dual-encoder framework employing contrastive learning to co-train a text encoder and a molecule encoder. The model aligns representations in a shared latent space and retrieves the most similar caption from the training set at inference time.
(2) **Generative Approach:** An encoder-decoder architecture where a Graph Neural Network (GNN) produces atom-level embeddings that condition a pretrained language model to generate novel captions autoregressively.
(3) **Hybrid Approach (MolCA-based):** A cross-modal framework [7] that combines a graph encoder with a Q-Former cross-modal projector and a large language model (Galactica-1.3B). This approach bridges the gap between retrieval and generation by translating molecular graph representations into soft prompts that the language model can understand, enabling open-ended text generation conditioned on 2D molecular structures.

Our experiments demonstrate that the hybrid MolCA-based approach significantly outperforms both alternatives, achieving a public leaderboard score of **0.846**, compared to 0.647 for the retrieval-based method and 0.52 for the generative approach. This result underscores the importance of explicit cross-modal alignment mechanisms combined with the generative capabilities of large pretrained language models.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Dataset

The dataset contains approximately 33,000 molecule–text pairs, with around 32,000 used for training and validation and 1,000 reserved for testing. Each molecule is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $v \in \mathcal{V}$ correspond to atoms and edges $e \in \mathcal{E}$ represent chemical bonds. Node features encode atomic properties (atom type, degree, formal charge, hybridization state), while edge features capture bond characteristics (bond type, conjugation, ring membership). The training data includes human-readable textual descriptions that characterize the molecular structure and chemical properties.

3 Text Encoders

For our retrieval-based approach, all text encoders are pretrained BERT-like models. We describe each variant below.

3.1 DistilBERT

DistilBERT [9] is a smaller and more efficient version of BERT [2], obtained through knowledge distillation. While reducing the model size by 40%, it retains 97% of BERT's language understanding capabilities and allows 60% faster inference. In the baseline implementation, DistilBERT is used to encode the natural language descriptions of molecules.

3.2 SciBERT

SciBERT [1] is a BERT-based model pretrained on a large corpus of scientific publications (1.14M papers) using a domain-specific vocabulary. Integrating SciBERT as a text encoder allows the model to better capture chemical and scientific terminology. However, due to its larger size and longer training time, initial experiments with SciBERT alone did not outperform DistilBERT when limited to a few training epochs.

3.3 RoBERTa

RoBERTa [6] is a robustly optimized BERT variant trained on large-scale general-purpose corpora. We initially compared RoBERTa with MPNet in preliminary experiments and found RoBERTa provided better overall textual embeddings for general domain content.

3.4 Combined Approach: SciBERT + RoBERTa

To leverage the strengths of both domain-specific and general-purpose language models, we implemented a dual-text encoder setup using SciBERT and RoBERTa. SciBERT provides in-depth understanding of chemical and scientific language, while RoBERTa captures broader contextual patterns. Empirically, this combination achieved the best performance for the retrieval-based approach,

outperforming individual encoders and the baseline DistilBERT model.

4 Graph Encoders

Graph Neural Networks (GNNs) serve as molecular encoders across all our approaches. We experimented with several architectures for the retrieval-based method, while the hybrid approach leverages a pretrained GINE encoder.

4.1 GATv2

To overcome the limitations of standard baseline GCNs, we used Graph Attention Networks (GATv2). Unlike GCNs, which aggregate neighboring node features using uniform weights, GATv2 applies an attention mechanism that allows the model to assign different importance to each neighboring atom during message passing. This is particularly relevant for molecular graphs, where certain bonds and atom types carry more chemical significance than others.

Formally, for a node v_i with neighbors $v_j \in \mathcal{N}(v_i)$, the attention coefficients are computed as:

$$\alpha_{ij}^{(t)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}^{(t)}\mathbf{h}_i^{(t)} \,\|\, \mathbf{W}^{(t)}\mathbf{h}_j^{(t)}\right]\right)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}^{(t)}\mathbf{h}_i^{(t)} \,\|\, \mathbf{W}^{(t)}\mathbf{h}_k^{(t)}\right]\right)\right)}, \tag{1}$$

where $\mathbf{h}_i^{(t)}$ denotes the representation of node v_i at layer t, $\mathbf{W}^{(t)}$ is a learnable weight matrix, $\mathbf{a}$ is a trainable attention vector, and $[\cdot\|\cdot]$ denotes concatenation. The node representations are then updated as:

$$\mathbf{h}_i^{(t+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(t)}\mathbf{W}^{(t)}\mathbf{h}_j^{(t)}\right), \tag{2}$$

where σ is a non-linear activation function.

4.2 Graph Isomorphism Network (GIN)

GIN [11] is designed to be maximally expressive among message-passing GNNs, approaching the discriminative power of the Weisfeiler-Lehman graph isomorphism test. At layer k, the representation of a node v is updated as:

$$\mathbf{h}_v^{(k)} = \text{MLP}^{(k)}\left((1+\epsilon^{(k)})\mathbf{h}_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} \mathbf{h}_u^{(k-1)}\right), \tag{3}$$

where $\mathbf{h}_v^{(k)}$ denotes the hidden state of node v at layer k, $\mathcal{N}(v)$ is the set of its neighbors, $\text{MLP}^{(k)}$ is a multilayer perceptron, and $\epsilon^{(k)}$ is a scalar parameter.

For our retrieval-based approach, we augment GIN with a virtual node that aggregates global graph information to improve information flow across distant parts of the molecular graph. For the hybrid MolCA-based approach, we use a five-layer GINE pretrained on 2 million molecules from the ZINC15 dataset using graph contrastive learning. Given a molecular graph g, this encoder produces node-level representations:

$$f(g) = \mathbf{Z} \in \mathbb{R}^{|g|\times d}, \tag{4}$$

where $|g|$ denotes the number of nodes and $d = 300$ is the hidden dimension.

4.3 GraphSAGE

GraphSAGE [3] is an inductive GNN framework based on a *Sample & Aggregate* strategy. For each node, a fixed-size subset of neighbors is sampled and their features are aggregated to update the node representation. The message-passing scheme is defined as:

$$\mathbf{m}_v^{(t)} = \text{AGGREGATE}\left(\{\mathbf{h}_u^{(t)} \mid u \in \mathcal{N}_k(v)\}\right), \tag{5}$$

$$\mathbf{h}_v^{(t+1)} = \sigma\left(\mathbf{W}^{(t)}\left[\mathbf{h}_v^{(t)} \,\|\, \mathbf{m}_v^{(t)}\right]\right). \tag{6}$$

5 Methods

We present our three distinct approaches to molecular graph captioning, each representing a different paradigm for cross-modal learning.

5.1 Retrieval-Based Approach

The retrieval-based approach employs a dual-encoder framework trained with contrastive learning to align molecular and textual representations in a shared latent space.

5.1.1 Contrastive Learning. We train our model using a contrastive learning objective to align molecular graph embeddings with their corresponding text descriptions. Given a batch of N molecule–text pairs, we compute normalized embeddings $\mathbf{v}_i^{\text{mol}}$ and $\mathbf{v}_i^{\text{text}}$ and define the similarity scores as:

$$s_{ij} = \frac{\mathbf{v}_i^{\text{mol}} \cdot \mathbf{v}_j^{\text{text}}}{\tau}, \tag{7}$$

where τ is a temperature hyperparameter. The contrastive loss is defined as a symmetric cross-entropy objective:

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\text{mol}\to\text{text}} + \mathcal{L}_{\text{text}\to\text{mol}}\right), \tag{8}$$

which encourages matching molecule–text pairs to have high similarity while pushing apart non-matching pairs within the batch.

Unlike the baseline MSE loss that directly minimizes the distance between each molecular embedding and its corresponding text embedding, the contrastive loss leverages all non-matching pairs in the batch as negatives, leading to a more discriminative embedding space.

5.1.2 Re-ranking Step. After the initial retrieval step based on cosine similarity, we apply a re-ranking strategy to refine the final selection of captions. We introduce a similarity score computed from graph-level chemical descriptors, including the number of atoms, edges, heteroatoms (oxygen and nitrogen), aromatic atoms, and graph density. For a given test molecule, we first retrieve the top-k candidates based on cosine similarity, then re-rank using:

$$\text{score} = \text{cosine} + \alpha \cdot \text{structure\_similarity}, \tag{9}$$

where α controls the influence of the structural component. This re-ranking step helps favor candidates that are not only semantically similar in the embedding space but also chemically consistent.

5.2 Generative Approach

In addition to the retrieval-based framework, we explored a fully generative approach to molecular description. Instead of selecting

an existing description from the training set, the goal was to generate a new textual description directly conditioned on the molecular graph.

5.2.1 Architecture. Our generative model follows an encoder-decoder mechanism. The molecular graph is first encoded using a GNN, producing a sequence of atom-level embeddings. These embeddings are then interpreted as encoder hidden states for a pretrained T5 model [8]. The decoder generates the molecular description autoregressively via cross-attention over the atom embeddings.

5.2.2 Limitations. Despite successfully training the generative model and achieving stable optimization, the resulting performance reached an accuracy of approximately 0.52, showing only marginal improvement over the baseline. Several factors explain this limited gain:

- **Capacity limitations:** The T5 model has limited expressive capacity for capturing complex chemical-linguistic mappings.
- **Weak graph-text alignment:** Without explicit alignment supervision, the cross-attention mechanism struggles to associate chemical substructures with the correct textual expressions.

5.3 Hybrid Approach: MolCA Model Fine-Tuning

Our best-performing approach is based on MolCA (Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter) [7], which addresses the limitations of both retrieval and purely generative methods.

5.3.1 MolCA Architecture Overview. MolCA enables a Large Language Model (LM) to understand both text-based and graph-based molecular content through three key components:

(1) **Graph Encoder:** A five-layer GINE pretrained on 2 million molecules produces node-level representations $\mathbf{Z} \in \mathbb{R}^{|g| \times d}$.
(2) **Cross-Modal Projector (Q-Former):** A Querying-Transformer [5] that bridges the graph encoder's representation space and the LM's text space. The Q-Former maintains $N_q = 8$ learnable query tokens $\{\mathbf{q}_k\}_{k=1}^{N_q}$ that interact with the graph encoder's output through cross-attention modules:

$$\mathbf{m}_k = \text{CrossAttn}(\mathbf{q}_k, \mathbf{Z}) = \text{softmax}\left(\frac{\mathbf{q}_k \mathbf{W}_Q (\mathbf{Z}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{Z}\mathbf{W}_V, \quad (10)$$

producing molecule representations $\{\mathbf{m}_k\}_{k=1}^{N_q}$ that capture text-relevant molecular features.
(3) **Language Model with LoRA:** Galactica-1.3B [10], a decoder-only transformer pretrained on scientific literature, generates captions conditioned on the soft prompts. LoRA (Low-Rank Adaptation) [4] enables efficient adaptation by adding trainable low-rank matrices to selected layers.

5.3.2 Why MolCA is a Hybrid Approach. MolCA fundamentally differs from both pure retrieval and pure generation approaches:

- Unlike **retrieval methods**, MolCA generates novel text rather than selecting from a fixed caption database, providing flexibility for describing unseen molecules.
- Unlike **pure generative methods** that directly map GNN outputs to a decoder, MolCA uses an explicit cross-modal projector (Q-Former) pretrained with multiple alignment objectives: molecule-text contrastive learning, molecule-text matching, and molecule captioning.
- MolCA leverages **dual molecular representations**: both 2D graphs (via the GINE encoder) and 1D SMILES strings (via the LM's tokenizer), combining structural and sequential information.

5.3.3 Cross-Modal Projection Layer. The critical interface between the Q-Former and the language model is the **cross-modal projection layer**. After the Q-Former extracts molecule features $\{\mathbf{m}_k\}_{k=1}^{N_q} \in \mathbb{R}^{d_q}$ through cross-attention with the graph encoder outputs, these must be mapped to the LM's embedding space. This projection is computed as:

$$\tilde{\mathbf{m}}_k = \mathbf{W}_{\text{proj}}\mathbf{m}_k + \mathbf{b}_{\text{proj}}, \quad (11)$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_{\text{LM}} \times d_q}$ and $\mathbf{b}_{\text{proj}} \in \mathbb{R}^{d_{\text{LM}}}$ are learnable parameters, with $d_{\text{LM}} = 2048$ for Galactica-1.3B.

The projected soft prompts $\{\tilde{\mathbf{m}}_k\}_{k=1}^{N_q}$ are then concatenated with the tokenized SMILES representation and fed to the language model for autoregressive generation.

5.3.4 Our Fine-Tuning Strategy. Starting from the pretrained MolCA checkpoint, we fine-tune on our challenge dataset. We systematically evaluated three fine-tuning configurations:

(1) **Full fine-tuning (GNN + Q-Former + LoRA):** Training the graph encoder, Q-Former, and LoRA adapters simultaneously. This configuration exhibited unstable training and suboptimal convergence due to catastrophic forgetting of pretrained knowledge.
(2) **LoRA-only fine-tuning:** Freezing the GNN and Q-Former while only updating LoRA parameters. This approach underfits the target domain because the cross-modal projector cannot adapt to dataset-specific patterns.
(3) **Cross-modal projection layer fine-tuning (selected):** Freezing the GNN, Q-Former attention weights, and LoRA adapters while training *only* the projection layer ($\mathbf{W}_{\text{proj}}, \mathbf{b}_{\text{proj}}$). This configuration achieved the best results.

Rationale: The pretrained components (GNN, Q-Former, LoRA) already encode rich molecular and linguistic knowledge from large-scale pretraining. The projection layer serves as the critical *bottleneck* for cross-modal transfer—adapting this interface to our dataset's specific molecule-text correspondences provides effective domain adaptation without disrupting the learned representations.

5.3.5 Training Objective. The model is trained using the standard language modeling loss:

$$\mathcal{L}_{\text{LM}} = -\sum_{l=1}^{L} \log p(y_l \mid y_1, \ldots, y_{l-1}, \{\tilde{\mathbf{m}}_k\}_{k=1}^{N_q}, \mathbf{s}), \quad (12)$$

where $\mathbf{y} = (y_1, \ldots, y_L)$ is the target description, $\{\tilde{\mathbf{m}}_k\}$ are the projected soft prompts, and $\mathbf{s}$ represents the tokenized SMILES input wrapped with special tokens `[START_I_SMILES]` and `[END_I_SMILES]`.

5.3.6 Training Details.

- **Learning rate:** 1×10^{-4} for projection parameters

- **Optimizer:** AdamW with weight decay 0.01
- **Scheduler:** Cosine annealing with warm restarts ($T_0 = 5$ epochs, $T_{\text{mult}} = 2$)
- **Batch size:** 10 with gradient accumulation (effective batch size 30)
- **Mixed precision:** BFloat16
- **Query tokens:** $N_q = 8$
- **Early stopping:** Patience of 10 epochs based on validation BLEU-4 and BERTScore
- **Beam search:** 5 beams for inference

5.3.7 Inference. At inference time, the model processes a molecular graph through the GINE encoder and Q-Former to produce soft prompts. These are concatenated with the tokenized SMILES and fed to Galactica, which generates the caption autoregressively using beam search ($B = 5$ beams):

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \prod_{l=1}^{L} p(y_l \mid y_{<l}, \{\tilde{\mathbf{m}}_k\}, \mathbf{s}). \tag{13}$$

6 Results

We evaluate all three approaches on the challenge dataset. The primary evaluation metric combines BLEU-4 and BERTScore to measure both n-gram overlap and semantic similarity between generated and reference descriptions.

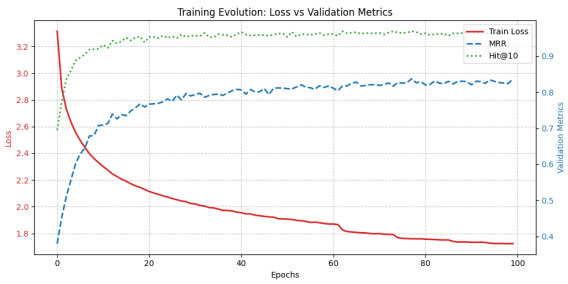

Figure 1: Training performance of the GIN model over 100 epochs, showing the evolution of training loss, MRR, and Hit@10

6.1 Analysis

The results demonstrate a clear performance hierarchy among the three paradigms:

Hybrid approach (0.846): The MolCA-based method achieves substantially higher performance by effectively combining molecular graph understanding with the generative capabilities of a large pretrained language model. The Q-Former's explicit cross-modal alignment mechanism—pretrained with contrastive, matching, and captioning objectives—enables the model to translate structural features into semantically meaningful soft prompts. Galactica's pretraining on scientific literature provides strong priors for generating chemically accurate descriptions.

Retrieval-based approach (0.647): The ensemble of GIN, GATv2, and GraphSAGE with SciBERT+RoBERTa text encoders achieves solid performance by leveraging contrastive learning for cross-modal alignment. However, this approach is fundamentally limited to selecting from existing training descriptions, which may not perfectly match novel test molecules.

Generative approach (0.52): The T5-based generative model shows limited improvement over the baseline. Without explicit cross-modal alignment mechanisms and with moderate model capacity, the direct encoder-decoder approach struggles to establish strong correspondences between atomic features and linguistic expressions.

7 Conclusion

In this work, we explored three paradigms for molecular graph captioning: retrieval-based, generative, and hybrid approaches. Our experiments demonstrate that the hybrid MolCA-based approach, which combines a pretrained GINE encoder with a Q-Former cross-modal projector and Galactica-1.3B language model, significantly outperforms both alternatives. Effective molecular captioning requires both strong cross-modal alignment and powerful language generation capabilities. The retrieval-based approach provides alignment but lacks generative flexibility, while the pure generative approach has generation capacity but weak alignment.

MolCA bridges this gap through its Q-Former architecture, which explicitly learns to translate molecular graph features into soft prompts via multi-task pretraining (molecule-text contrastive learning, matching, and captioning).

Our fine-tuning strategy—adapting only the cross-modal projection layer while keeping the pretrained GNN, Q-Former attention, and LoRA adapters frozen—proved most effective. This suggests that the pretrained components already capture rich molecular and linguistic knowledge, and the projection interface is the critical bottleneck for domain adaptation. The dual molecular representation (2D graphs + 1D SMILES) in MolCA provides complementary structural and sequential information that enhances caption quality.

Future work: Future work could explore: (1) incorporating 3D molecular geometry for more complete structural understanding, (2) multi-task learning with related objectives such as property prediction, and (3) scaling to larger language models with more extensive molecular pretraining data.

The significant performance gap between our hybrid approach (0.846) and the alternatives (0.647 and 0.52) underscores the importance of explicit cross-modal alignment mechanisms in molecular language modeling tasks.

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv:1903.10676 [cs.CL] https://arxiv.org/abs/1903.10676

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[3] William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive Representation Learning on Large Graphs. arXiv:1706.02216 [cs.SI] https://arxiv.org/abs/1706.02216

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV] https://arxiv.org/abs/2301.12597

	Number of Epochs					
Graph Encoder	**30**	**60**	**90**	**140**	**190**	**Best MRR**
GCN (Baseline)	0.248	0.488	-	-	-	0.488
GATv2	0.275	0.421	0.507	-	-	0.507
GIN	0.318	0.473	0.555	0.590	0.608	0.608
GraphSAGE	0.243	0.467	0.497	0.515	0.551	0.551

Table 1: MRR on validation set for individual graph encoders in the retrieval-based approach. Results show the evolution of performance across training epochs with SciBERT+RoBERTa text encoders.

Approach	**Model Configuration**	**Public Score**
Retrieval-Based	GIN + GraphSAGE (Ensemble)	0.640
	GIN + GATv2 + GraphSAGE (Ensemble)	0.647
Generative	T5 + GNN Encoder	0.52
Hybrid (MolCA)	**GINE + Q-Former + Galactica-1.3B**	**0.846**

Table 2: Comparison of all three approaches on the public leaderboard. The hybrid MolCA-based approach with cross-modal projection fine-tuning significantly outperforms both retrieval-based ensembles and the generative model.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[7] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2024. MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter. arXiv:2310.12798 [cs.CL] https://arxiv.org/abs/2310.12798

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG] https://arxiv.org/abs/1910.10683

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL] https://arxiv.org/abs/1910.01108

[10] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. arXiv:2211.09085 [cs.CL] https://arxiv.org/abs/2211.09085

[11] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=ryGs6iA5Km arXiv:1810.00826.