Named Entity Recognition of Thyroid Diseases Based on the BioBERT Pre-trained Model

Yanghui Huang[2], Guiling Guo[3], Huiqing Jia[2], Chengjie Luo[2], Zeyu Liu[1]

[1]Institute of Artificial Intelligence
[2]School of Informatics
[3]School of Medicine
31520241154506, 24520240157767, 31520241154507, 31520241154490, 36920241153241

Abstract

Biomedical text mining plays a crucial role in extracting valuable information from the growing volume of biomedical literature. However, applying general NLP advancements directly to biomedical text often results in suboptimal performance due to domain-specific challenges. This study explores the effectiveness of advanced graph-based approaches and pre-trained language models in addressing these challenges, with a focus on thyroid disease analysis.

We propose a novel framework combining Named Entity Recognition (NER) and relation extraction techniques with a knowledge graph powered by Graph Attention Networks (GATs). The framework effectively identifies key medical entities and relationships while enhancing the completeness of the knowledge graph through link prediction techniques. Leveraging domain-specific embeddings generated by BioBERT, the proposed method demonstrated significant improvements in capturing complex medical relationships compared to baseline models.

Our experiments highlight the model's superiority in accuracy and robustness, with a detailed case study illustrating its practical utility in identifying nuanced medical connections. This work underscores the potential of integrating GNNs and pre-trained models to advance medical text mining and decision support systems.

Introduction

With the rapid advancement of biomedical research, the volume of biomedical literature is experiencing explosive growth. These literatures contain a wealth of valuable information about new discoveries and insights, which are crucial for advancing medical progress and improving patient care. However, due to the unstructured nature of this textual data, extracting useful information from it poses significant challenges. Against this backdrop, biomedical text mining has become increasingly important, aiming to automatically identify and extract key information such as disease names, medications, genes, and proteins from the literature.

Named Entity Recognition (NER) is a core task in biomedical text mining, involving the identification and classification of specific entities within text, such as personal names, locations, organizations, and domain-specific terminology. The development of NER techniques has far-reaching implications for information retrieval, clinical decision support systems, drug discovery, disease monitoring, and more. Traditional NER methods, such as rule-based systems and classical machine learning models, have achieved certain successes within their respective domains. However, they often fall short when dealing with the complexity and specificity of biomedical text(Ahmad, Shah, and Lee 2023).

In recent years, the advent of deep learning techniques, especially pre-trained language models, has significantly improved the performance of NER tasks. BioBERT, a pre-trained model specifically designed for the biomedical domain, has demonstrated substantial improvements in understanding and entity recognition in biomedical text by pre-training on large-scale biomedical corpora (Lee et al. 2020). BioBERT has shown superior performance in a variety of biomedical text mining tasks, including biomedical named entity recognition, relation extraction, and question answering(Lee et al. 2020).

Thyroid diseases, as a common category of endocrine system disorders, are of great importance for enhancing the quality of life of patients. However, due to the vast number of thyroid disease-related biomedical literatures and the specialized terminology involved, traditional NER methods often fail to deliver satisfactory results. Therefore, developing and applying more advanced NER technologies, especially pre-trained models like BioBERT, is of significant importance for improving the efficiency and accuracy of information extraction in thyroid disease-related literature.

This study aims to explore the efficacy of using the BioBERT pre-trained model for entity extraction in texts related to thyroid diseases. We believe that by leveraging the capabilities of BioBERT, we can more accurately identify disease-related entities in the text, thereby providing a richer information resource for the research and clinical application of thyroid diseases.

Related Work

Named entity recognition (NER) is a cornerstone task within the field of natural language processing (NLP), particularly significant in the biomedical domain due to its potential to enhance clinical decision support systems and real-world medical research. The historical progression of NER has seen a shift from rule-based and dictionary-based methods,

which are limited by the need for extensive expert knowledge and scalability issues, to statistical machine learning approaches. More contemporary methods have embraced deep learning, which has demonstrated an exceptional ability to generalize and extract features from large datasets.

Deep neural networks (DNNs), and more specifically, recurrent neural networks (RNNs) such as bidirectional long short-term memory (BiLSTM) networks, have become prevalent due to their efficacy in capturing sequential dependencies within text. The synergy between BiLSTM and conditional random fields (CRFs) has emerged as a dominant model in NER, effectively addressing both local context features and the sequential constraints of label predictions(Fu et al. 2024).

The incorporation of deep learning has led to the exploration of various network architectures and enhancements. Collobert et al. (2011) were pioneers in applying deep learning to NER across multiple information extraction tasks, showcasing the potential of neural networks in feature learning. Socher (2012) introduced the MV-RNN model, harnessing word vectors to understand the relationships between words, thus advancing the field of NER. Attention mechanisms have since been integrated into neural networks to focus on local context, further improving feature extraction for NER.

In the context of biomedical NER, the complexity of medical texts, replete with nested entities, abbreviations, and ambiguous terms, presents unique challenges. To overcome these, researchers have proposed models that capitalize on domain-specific knowledge and advanced neural network architectures. Gao et al. (2019) proposed an attention-based ID-CNNs-CRF model to enhance NER performance by combining word-order features with local context. Chowdhury (2018) developed a multi-task RNN model aimed at extracting medical entities from Chinese electronic medical records, demonstrating the applicability of deep learning in cross-linguistic NER tasks(Lai and Jie 2023).

The introduction of pre-trained language models, such as BERT, has marked a significant advancement in NER by providing deep contextualized word representations. BERT's capacity to capture bidirectional context has been highly beneficial for NER tasks. Subsequent models have integrated BERT with other neural network architectures, like BiLSTM and CRF, to further bolster NER performance. Mou (2022) designed a medical encoding transformer that leverages BERT for semantic enhancement in named entity recognition, showcasing the integration of pre-trained models with advanced neural network architectures for improved NER.

Despite these advancements, accurately recognizing entities in electronic medical records (EMRs) remains challenging, often due to the intricacy of language structure and the scarcity of large-scale annotated datasets. To enhance NER accuracy, researchers have investigated data augmentation techniques and model fusion strategies. For instance, a method proposed by Xuewei Lai et al. (2023) combines BERT's semantic enhancement with BiLSTM and CRF to address the lack of local context features and improve entity recognition accuracy(Wang et al. 2023).

Researchers like Zhu et al. (2017) have explored online medical prediagnosis frameworks with high efficiency and privacy protection, utilizing nonlinear kernel SVMs, while Wang et al. (2018) focused on memory mechanisms and proposed gradient-based learning algorithms to address classification tasks.(Liu et al. 2021) These works highlight the multifaceted approach towards enhancing NER in the medical domain.

In summary, the related work in NER encompasses a spectrum of approaches, from traditional machine learning to state-of-the-art deep learning models. There is a pronounced trend towards integrating pre-trained language models like BERT with neural network architectures to enhance feature extraction and improve recognition accuracy within the biomedical domain. The continuous evolution of NER models, as evidenced by the works of Collobert et al. (2011), Socher (2012), Gao et al. (2019), Chowdhury (2018), and Mou (2022), reflects the dynamic nature of the field and the ongoing pursuit of more accurate and context-aware NER solutions.

Methodology

We will construct a medical knowledge graph specifically focused on thyroid cancer and utilize link prediction techniques to identify and fill in the missing relationships within the graph.This entails extracting relevant entities and relationships from a vast corpus of medical literature, analyzing and inferring potential connections between entities using advanced graph neural network models, thereby enhancing the completeness and accuracy of the knowledge graph. The knowledge graph is designed to support research and clinical decision-making in the field of thyroid cancer by revealing complex relationships between disease symptoms, treatment methods, and research findings, providing medical professionals with a powerful data-driven tool.

Data Collection and Preprocessing

The construction of a knowledge graph for thyroid cancer is impeded by the absence of a comprehensive, pre-existing dataset that encompasses the breadth of information related to thyroid diseases and, more specifically, thyroid cancer. This gap in data availability underscores the necessity of developing a custom dataset tailored to the intricate requirements of knowledge graph construction. The dataset must be rich in detail, covering a wide array of aspects such as disease symptoms, treatment modalities, and the latest research findings to ensure its utility in both research and clinical applications.

Data Collection: The cornerstone of our dataset is the meticulous collection of textual data from reputable medical sources. To assemble a robust corpus that serves as a solid foundation for our knowledge graph, we have adopted a strategic approach to data collection. Our focus is on scholarly articles that are directly pertinent to thyroid diseases, with a particular emphasis on thyroid cancer. The keyword "Thyroid cancer" has been identified as the primary search term to guide our collection efforts. We have chosen PubMed, a premier database of biomedical literature, as

our primary source of articles. Our collection strategy is to extract the 1,000 most-cited articles on thyroid cancer from PubMed, as these articles are likely to contain seminal research and widely accepted findings. These texts will not only provide a comprehensive overview of the disease but also offer insights into various symptoms, treatment strategies, and significant research breakthroughs, thereby forming the backbone of our knowledge graph.

Text Preprocessing: Once the textual data has been collected, the next critical phase is text preprocessing. This step is indispensable for preparing the data for subsequent analysis and knowledge graph construction. The preprocessing involves several sub-steps, each designed to refine the data and enhance its quality. Initially, we will remove any irrelevant content that does not contribute to the understanding of thyroid cancer. This may include extraneous information, redundant text, or sections that are not pertinent to our study. Following this, we will tokenize the text, breaking it down into its constituent parts such as words, phrases, or symbols. This process facilitates the identification and extraction of key entities and relationships within the text(Xue et al. 2019). Additionally, we will standardize medical terminologies to ensure consistency across the dataset. Medical language is often complex and varied, with different terms used interchangeably. Standardization is crucial to avoid ambiguity and to ensure that the entities extracted are both accurate and consistent. This preprocessing step is fundamental to the integrity of our knowledge graph, as it directly impacts the quality and reliability of the information it contains.

Named Entity Recognition Task

In the realm of medical text analysis, the Named Entity Recognition (NER) task is pivotal for identifying and categorizing key medical entities such as diseases, drugs, and symptoms. This task is essential for constructing a comprehensive and accurate knowledge graph that can aid in medical research and clinical decision-making.

NER Model Selection: For the identification of key medical entities, we have chosen to employ the Llama 3.1-8b-Instruct model in conjunction with GraphRAG(Edge et al. 2024). GraphRAG, developed by Microsoft, is an innovative technique designed to enhance RAG (retrieval-augmented generation) systems. It leverages knowledge graphs to contextualize and improve the information retrieval process. GraphRAG addresses the challenges faced by traditional RAG systems in handling complex queries and multi-hop reasoning by integrating a knowledge graph memory structure. This integration allows for more accurate and contextually relevant information retrieval, making it an ideal choice for our NER task.

Entity and Relationship Extraction: Following the identification of entities, the next step is to extract relationships from the text, such as "Disease-Symptom" or "Drug-Treatment". This involves a detailed annotation process where entities are categorized into different classes, including "Disease", "Drug", "Event", and others. These annotations are crucial for building a knowledge graph that accurately represents the interconnections between various medical entities(Dong et al. 2014). The extraction process will involve sophisticated natural language processing techniques to ensure that the relationships are accurately captured and represented.

Visualization: To make the complex structure of the knowledge graph more accessible and understandable, we will employ graph-based visualization tools. These tools will help us to visually represent the knowledge graph, offering insights into the interrelationships between various medical entities(Suchanek, Kasneci, and Weikum 2007). Visualization is a powerful method to communicate the structure and significance of the relationships within the graph, making it easier for researchers and clinicians to interpret and utilize the information.

The outcome of this process will be a foundational NER and RE (relation extraction) result that is meticulously annotated. The annotations will include categories such as "Disease", "Drug", "Event", and more.

By following this expanded approach, we aim to create a robust and informative knowledge graph that can significantly contribute to the field of medical research and practice.

Graph Construction and Link Prediction

Word Embeddings: In the realm of graph construction, incorporating semantic understanding is crucial for capturing the essence of relationships within the graph. To achieve this, we will utilize BioBERT(Lee et al. 2020), a domain-specific pre-trained language model that has been fine-tuned for biomedical text. BioBERT's capabilities in generating word embeddings will be harnessed to infuse our graph with a deeper semantic comprehension. These embeddings will serve as the foundation for understanding the nuanced relationships between medical terms, thereby enhancing the graph's ability to represent complex medical concepts accurately.

Link Prediction Model: A pivotal aspect of graph construction is the prediction of missing relationships between entities, which can significantly enhance the graph's completeness and utility. To address this, we will develop and implement a sophisticated link prediction model. For comparison purposes, we will use StellarGraph as a baseline(Data61 2018). StellarGraph, a library designed for graph machine learning, provides a standard approach to link prediction.

Comparison with GAT Model: After establishing the baseline using StellarGraph, the next step will be to compare its performance with that of a Graph Attention Network (GAT) model(Velickovic et al. 2017). GAT models have gained recognition for their attention-based mechanisms, which are particularly adept at handling complex graph structures. By incorporating attention mechanisms, GAT models can prioritize certain relationships over others, leading to more accurate and efficient link predictions. This comparison will be crucial in evaluating the effectiveness of different graph neural network architectures in the context of medical knowledge graph construction. We anticipate that the GAT model will offer significant improvements in performance, particularly in scenarios where the graph's

structure is intricate and the relationships between entities are multifaceted.

Graph Neural Network (GNN) Construction

We will utilize a GAT (Graph Attention Network) for the final knowledge graph construction. This model, particularly suited for relational data like knowledge graphs, will allow us to leverage attention mechanisms to capture the most relevant relationships between entities. GATs are a variation of Graph Convolutional Networks (GCNs) that incorporate the concept of attention mechanisms, allowing nodes to focus on the characteristics of their neighborhoods without having to perform an expensive matrix operation (like inversion) or rely on prior knowledge of the graph's structure.The following pseudocode outlines the core components of the GNN architecture and its training process:

Algorithm 1: GNN Architecture and Training Process

Input: Input features x, edge indices $edge\_index$, training data $train\_loader$

Output: Trained GNN model, loss history

Class GNN:
Initialize: Define 3 GATConv layers with specified input/output dimensions; define a dropout layer;
Forward: Apply GATConv layer 1 $\rightarrow$ ReLU $\rightarrow$ dropout $\rightarrow$ GATConv layer 2 $\rightarrow$ ReLU $\rightarrow$ dropout $\rightarrow$ GATConv layer 3 $\rightarrow$ output

Class Classifier:
Forward: Compute dot product of x_source and x_target $\rightarrow$ sigmoid function $\rightarrow$ probability

Class Model:
Initialize: Create instances of GNN and Classifier;
Forward: Pass data through GNN to get node embeddings $\rightarrow$ use Classifier to compute similarity between node pairs $\rightarrow$ predictions

Function train_model:
Initialize: Set device to GPU (if available) or CPU; move model to device; create Adam optimizer; initialize loss history;
for *each epoch from 1 to 100* **do**
  Initialize $total\_loss \leftarrow 0$, $total\_examples \leftarrow 0$ **for** *each batch in* $train\_loader$ **do**
    Zero out gradients Move batch data to device Get predictions from model Get ground truth labels Compute binary cross-entropy loss Perform backpropagation Update model parameters Accumulate $total\_loss$ and $total\_examples$
  if *epoch is a multiple of 10* **then**
    Print current loss

return trained model and loss history

GNN Optimization: The performance of the GAT model will be optimized through hyperparameter tuning and training on the annotated graph data. The model will be trained to predict links (relationships) between nodes (entities) within the graph. Optimization techniques such as neighborhood sampling and vertex mini-batching have been proposed to cope with GPUs' limited memory capacity, but these workarounds have drawbacks. They induce vast underutilization of the compute capacity, may degrade the network accuracy, and induce significant overhead through additional costly operations. Instead, a CPU can work with full-batches without sampling for large graphs, enabling wider and deeper network structures. Additionally, the GAT model supports representation learning and node classification for homogeneous graphs, with versions of the graph attention layer that support both sparse and dense adjacency matrices(Tong, Li, and Liu 2024).

Attention Mechanisms in GNN: The introduction of attention mechanisms in GNNs, which have been brilliant in the fields of natural language processing and computer vision, allows for adaptively selecting discriminative features and automatically filtering noisy information. This has led to significant advances in attention-based GNNs, which are surveyed comprehensively in recent literature(Ying et al. 2021).

Fine-Tuning with Large Language Models

Fine-tuning large language models enhances their accuracy and adaptability in specific domains while reducing training costs and data requirements. This approach enables models to handle domain-specific tasks more effectively and provides personalized solutions for research and applications.

Dataset Preparation for Fine-Tuning: To improve predictions, we will prepare a fine-tuning dataset that pairs knowledge graph triples (e.g., [Disease, Symptom, Drug]) with textual descriptions. This enables large language models (LLMs) to understand the context of entities and relationships within the medical domain(Parthasarathy et al. 2024). The dataset will be carefully curated to include a wide range of medical concepts and their interrelations, ensuring that the LLMs can learn from a comprehensive set of examples. This preparation is crucial for the model to grasp the nuances of medical terminology and the specific contexts in which these terms are used.

Fine-Tuning Process: We will fine-tune a pre-trained LLM (e.g., GPT-based models) on the dataset. This will enhance the model's ability to predict new relationships or answer domain-specific queries. The LLM will generate new insights or suggest potential treatment strategies based on the knowledge graph's context, thus contributing to the discovery of new therapeutic avenues.

Experiments

Knowledge Graph Construction

The knowledge graph was constructed using the methods described in the "Data Collection and Preprocessing" subsection of the Methodology section. The knowledge graph consists of nodes representing key medical entities such as diseases, symptoms, treatments, and relationships between these entities, such as "Disease-Symptom" and "Drug-Treatment." To illustrate the structure of the knowl-

edge graph, a partial visualization is included in Figure 1. The graph includes the following statistics:

Node Data:

- Total nodes: 924
- Unique node IDs: 308
- Duplicate node IDs: 616

Edge Data:

- Total edges: 326
- Unique source nodes: 102
- Unique target nodes: 247

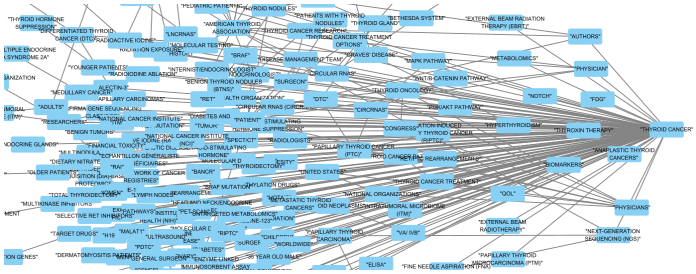

Figure 1: Graph Structure Example

Model Comparison

To evaluate the effectiveness of the proposed approach, we conducted comparative experiments against a baseline model, StellarGraph, using standard evaluation metrics such as AUC. Table 1 shows the performance comparison:

Model	AUC
StellarGraph	0.5781
Proposed Model	0.6527

Table 1: AUC scores of the baseline and the proposed model.

The results demonstrate that the proposed model significantly outperforms StellarGraph, showcasing the effectiveness of integrating advanced graph attention mechanisms into the knowledge graph construction process.

Case Analysis

To assess the practical utility of the model, we conducted a detailed case analysis. One illustrative example involves the relationship between familial non-medullary thyroid carcinoma (FNMTC) and second malignant neoplasm (SMN). The baseline model, StellarGraph, erroneously interpreted SMN as a gene, leading to outputs that overly emphasized genetic mutations. By contrast, the proposed model correctly captured the relationship and stated:

> "FNMTC has a higher risk of SMN compared to sporadic NMTC."

This result aligns with domain-specific medical knowledge and highlights the robustness of the proposed model in understanding nuanced medical relationships. Such improvements are critical for constructing clinically relevant and reliable knowledge graphs.

Conclusion

Summary

This study presents an innovative approach to constructing and utilizing a knowledge graph for thyroid disease diagnosis and analysis, incorporating advanced Graph Attention Network (GAT) mechanisms. The proposed model demonstrated a significant improvement over the baseline (StellarGraph) in terms of performance metrics such as AUC, underscoring its effectiveness in capturing and representing complex relationships within medical data.

Moreover, detailed case analyses revealed the model's robustness in understanding nuanced medical relationships that traditional methods often misinterpret. The constructed knowledge graph, enriched with detailed structures and inter-entity relationships, provides a powerful resource for understanding thyroid disease and its associated conditions. It not only supports clinical decision-making but also facilitates further research by uncovering previously unrecognized connections between entities.

Recommendations and Further Considerations

- **Data Diversity and Quality**
  While PubMed is an excellent resource, it is essential to ensure the diversity and quality of the data set. Incorporating additional sources, such as clinical trial databases or medical textbooks, may help broaden the knowledge base and fill potential gaps.
- **Entity Resolution**
  When performing Named Entity Recognition (NER), attention should be given to entity disambiguation, especially in cases where similar terms may refer to different concepts (e.g., "Thyroid cancer" vs. "Cancer of the thyroid"). Using advanced techniques such as entity linking can help resolve these ambiguities.
- **Scalability and Efficiency**
  As the project progresses, it is important to consider the scalability of the system. Fine-tuning LLMs and training GNNs can be resource-intensive, so optimizing the computational efficiency will be critical for handling large datasets.
- **Evaluation Metrics**
  Throughout the project, collaboration with medical domain experts is crucial to ensure the clinical relevance of the extracted entities and relationships, as well as the validity of the knowledge graph.
- **Collaboration with Domain Experts**
  Engage with medical experts to ensure the clinical relevance and validity of the extracted entities, relationships, and the knowledge graph.

By following this pipeline and incorporating these recommendations, the project has the potential to make significant contributions to medical knowledge discovery, particularly in the domain of thyroid cancer research.

References

Ahmad, P. N.; Shah, A. M.; and Lee, K. 2023. A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. In *Healthcare*, volume 11, 1268. MDPI.

Data61, C. 2018. StellarGraph Machine Learning Library. https://github.com/stellargraph/stellargraph.

Dong, X.; Gabrilovich, E.; Heitz, G.; Horn, W.; Lao, N.; Murphy, K.; Strohmann, T.; Sun, S.; and Zhang, W. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 601–610.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Fu, L.; Weng, Z.; Zhang, J.; Xie, H.; and Cao, Y. 2024. MMBERT: a unified framework for biomedical named entity recognition. *Medical & Biological Engineering & Computing*, 62(1): 327–341.

Lai, X.; and Jie, Q. 2023. A Named Entity Recognition Approach for Electronic Medical Records Using BERT Semantic Enhancement and BiLSTM. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 19(1): 1–14.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Liu, N.; Hu, Q.; Xu, H.; Xu, X.; and Chen, M. 2021. MedBERT: A pretraining framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics*, 18(8): 5600–5608.

Parthasarathy, V. B.; Zafar, A.; Khan, A.; and Shahid, A. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*.

Suchanek, F. M.; Kasneci, G.; and Weikum, G. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, 697–706.

Tong, G.; Li, D.; and Liu, X. 2024. An improved model combining knowledge graph and GCN for PLM knowledge recommendation. *Soft Computing*, 28(6): 5557–5575.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wang, T.; Zhang, Y.; Zhang, Y.; Lu, H.; Yu, B.; Peng, S.; Ma, Y.; and Li, D. 2023. A hybrid model based on deep convolutional network for medical named entity recognition. *Journal of Electrical and Computer Engineering*, 2023(1): 8969144.

Xue, K.; Zhou, Y.; Ma, Z.; Ruan, T.; Zhang, H.; and He, P. 2019. Fine-tuning BERT for joint entity and relation extraction in Chinese medical text. In *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 892–897. IEEE.

Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T.-Y. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34: 28877–28888.