A Zero-shot LLM-based Framework for Descriptive Gene-gene Interaction Network Generation

Bingjun Li [1] **Mason Zito Ritchotte** [1] **Sihong He** [2] **Sheida Nabavi** [1]

Abstract

The existing graph-based analytic methods for single-cell RNA sequencing utilize gene-gene interaction networks. However, existing gene-gene interaction databases like STRING and BioGrid have significant drawbacks, including incompleteness, static nature, and lack of descriptive information. Recently, large language model (LLM) has demonstrated powerful capabilities in understanding and reasoning in the biomedical field without any fine-tuning. To address the drawbacks of traditional gene-gene interaction databases and use the reasoning capabilities of LLMs, we propose LLM-GeneGraph, a novel generative framework that can dynamically generates detailed, scientifically validated gene-gene interaction networks by integrating retrieval-augmented generation (RAG) techniques and employing an ensemble of three state-of-the-art LLMs along with an LLM-as-a-judge. The proposed method shows its capabilities to produce validated interactions and novel insights beyond the scope of traditional databases.

1. Introduction

Advances in single-cell RNA sequencing (scRNA-seq) make profiling millions of cells at near-transcriptome scale more accessible to biologists and generate large quantities of sequencing data. To effectively analyze those complex high-dimensional datasets, researchers have developed many computational methods (Li & Nabavi, 2024; Zhao et al., 2022; Lin et al., 2022b; Li & Nabavi, 2023). Many of these approaches utilize the graph neural network by incorporating gene-gene interaction networks as prior knowledge (Li & Nabavi, 2023; Lin et al., 2022b). The quality and comprehensiveness of these gene-gene interaction networks significantly impact downstream analyses.

The two most widely adopted databases for gene-gene interactions are STRING(Szklarczyk et al., 2023) and BioGrid(Oughtred et al., 2021). These databases computed the interaction networks by mining publicly available literature and gene expression data. However, they suffer from three major drawbacks: (i) incompleteness, these gene-gene interaction networks are biased toward well-studied pathways; (ii) static nature, these databases are expensive to maintain and updated infrequently; and (iii) lack of information, as interactions are represented by scalar scores without any descriptive text about the interaction mechanism. We aim to propose a novel interaction generation framework to address all these drawbacks.

Recent development in large language models (LLM) has led to many novel methods for text reasoning and generation, demonstrating remarkable performance (Wei et al., 2022; Achiam et al., 2023; Liu et al., 2024). LLMs have also shown a strong understanding and reasoning within the biomedical domain without any fine-tuning (Chen & Zou, 2024). Furthermore, researchers have expanded LLM's strong text reasoning capability to graph-structured data (Wang et al., 2024; He et al., 2024; Li et al., 2024). Among these works, textual graphs get a lot of attention due to their rich information and real-world applications (Zhu et al., 2024; Tianxiang Jin et al., 2023). However, this powerful textual graph approach is currently unable to be applied to genomic analysis due to the drawbacks of current gene-gene interaction networks. Inspired by LLM's powerful understanding of genomic and textual graphs, we aim to address the following question:

Can we utilize LLMs to create a new descriptive gene-gene interaction network that is scientifically accurate, biologically validated, and provides new knowledge about genomics beyond the current networks?

To address this challenge, we propose LLM-GeneGraph, the first generative framework that utilizes zero-shot LLM's powerful understanding of genomics. LLM-GeneGraph is able to be updated dynamically and produce rich information of the interaction mechanism. Our contributions are: (i) a novel LLM-based generative framework for descrip-

[1]School of Computing, University of Connecticut, Storrs, CT, USA [2]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA. Correspondence to: Sheida Nabavi <sheida.nabavi@uconn.edu>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

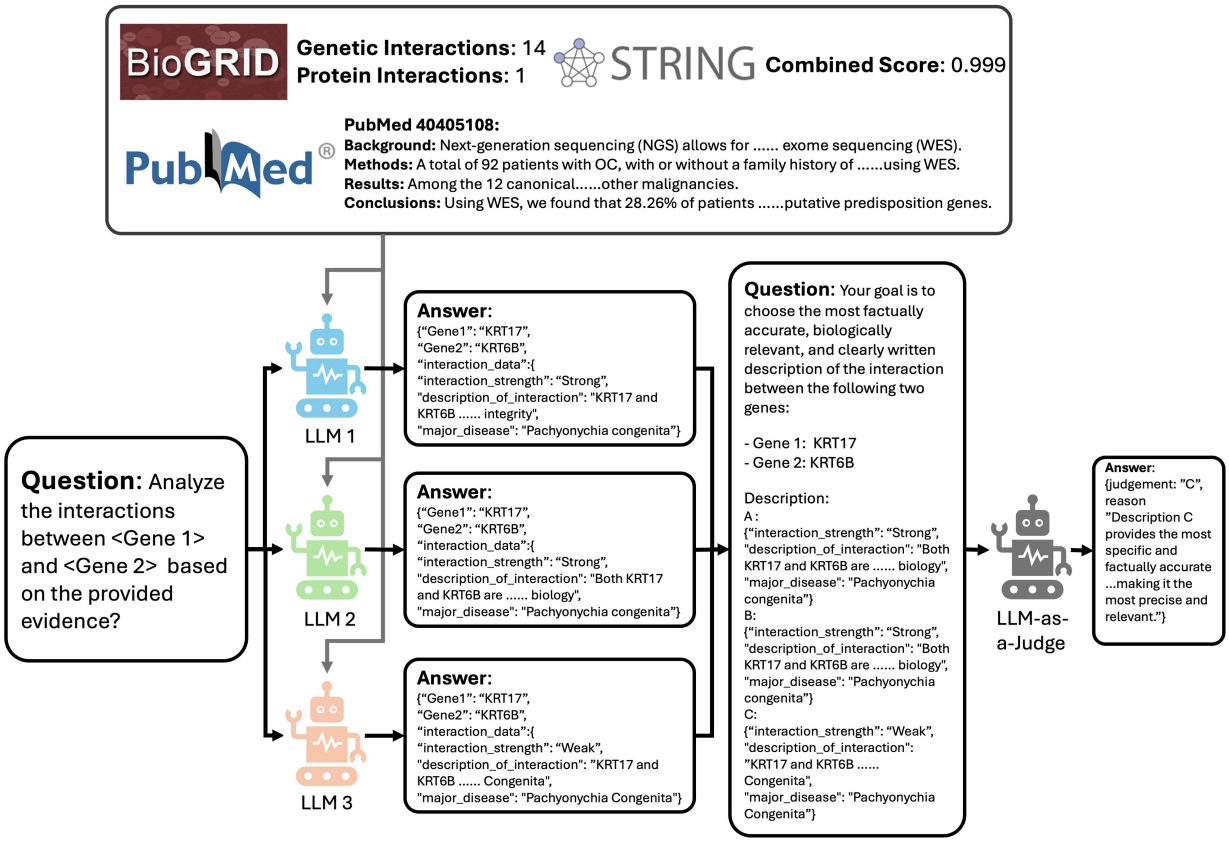

Figure 1. The proposed Gene-gene interaction network generation framework is shown. It consists of three major component: (a) the retreieval-agumented generation (RAG) that passes all the known information about the gene pair to the LLM model, which includes BioGrid database, STRING database and the abstracts of top 3 papers resulted by searching this gene pair on PubMed; (b) three independent LLMs that take the prior information and generate the description of the gene pair interactions; (c) a judge based on a reasoning LLM that takes all three descriptions and ranks the best one out of factual accuracy, biological relevance, and clear writing.

tive gene-gene interaction networks; (ii) an ensemble structure integrating three state-of-the-art LLM with a reasoning LLM as a judge for candidate selection; (iii) comprehensive validation, including scientific soundness checks against current genetic interaction networks and biological validation of newly generated interactions.

2. Related Works

As mentioned earlier, traditional gene-gene interaction databases, like BioGrid and STRING databases, suffer from several major drawbacks, such as incompleteness, static nature, and lack of information. These drawbacks can severally deteriorate the performance of downstream tasks based on these graphs. Bertin et al. show that even well-curated graphs often fall short of capturing the full dependencies observed in gene expression datasets, especially in single-cell contexts where interactions can vary dramatically by cell state or tissue (Bertin et al., 2019).

There are several studies that leverage LLMs to generate, enhance, or validate biomedical knowledge graphs (Rosenbaum et al., 2024; Feng et al., 2025). MedG-KRP is a graph-based probe to validate and interpret LLM's biomedical reasoning by comparing generated knowledge graphs to curated ontologies like BIOS (Rosenbaum et al., 2024). Knowledge Graph-based Thought (KGT) demonstrates that LLMs can interface with biomedical graphs to significantly reduce hallucinations and improve factual accuracy in pan-cancer question (Feng et al., 2025). These approaches demonstrate LLMs' potential to serve as dynamic and context-aware knowledge graph generators or validators.

To the best of our knowledge, LLM-GeneGraph is the first to utilize LLMs to generate descriptive gene-gene interaction network along with an LLM-as-a-judge. Our work aims to address all the drawbacks of traditional gene-gene interaction databases and provides a foundation for future single-cell analysis tools to built upon.

3. Method

As shown in Figure 1, the proposed interaction generation framework has three major components: (i) the retrieval-augmented generation (RAG) module; (ii) the ensemble

Table 1. A Demo of the Gene-gene Interaction Network Generated by LLM-GeneGraph Vs. BioGrid and STRING Databases

Gene Pair	BioGrid	STRING	LLM-GeneGraph
KRT17, KRT6B	1.5	0.874	**Strength:** Strong **Description:** KRT17 and KRT6B are both keratin proteins often co-expressed in epithelial tissues. Their interaction is important in the structural integrity of skin and epithelial cells, and they are known to participate in the pathogenesis of skin disorders such as Pachyonychia Congenita. **Disease:** Pachyonychia Congenita
IL8, WNT2	0	0	**Strength:** Weak **Description:** Based on the provided PubMed abstracts, IL8 and WNT2 are both expressed in gastric cancer and their expression levels are associated with the prognosis of the disease. However, there is no direct evidence of a strong interaction between these two genes from the STRING or BioGRID scores or the abstracts. The abstracts suggest a potential correlation in expression in a specific context (gastric cancer), but not necessarily a direct interaction. **Disease:** Gastric Cancer

structure of three LLMs; (iii) a reasoning LLM as the judge.

3.1. Retrieval-Augmented Generation (RAG)

The major challenge in generating scientifically accurate content with LLMs is hallucinations. Previous studies have shown that RAG can help reduce hallucinations and improve the factual accuracy of LLM outputs (Shuster et al., 2021; Ayala & Bechard, 2024). We integrate both BioGrid and STRING databases, and the relevant PubMed literature of each gene pair to provide context for gene–gene interaction generation. The genetic and protein interaction counts were obtained from BioGRID (Oughtred et al., 2021). The genetic interactions have 1.0 weight and the protein interactions have 1.5 weight, and the total weight is log normalized across all human genetic interactions. The confidence scores for the gene-gene interactions were retrieved from STRING data base (Szklarczyk et al., 2023). Both BioGrid and STRING scores are scaled between 0 and 1. The relevant PubMed literature was the abstracts of the top three search results by gene names. All three types of RAG information are passed into the LLM models in a JSON format as the evidence basis for assessing and generating detailed descriptions of gene-gene interactions.

3.2. Zero-Shot Interaction Generation

Previous study has shown that LLMs have been trained on sufficient biomedical literature and have significant capabilities to understand, reason, and generate biomedical texts (Chen & Zou, 2024; Ayers et al., 2023). We found that given identical RAG information and question prompts, different LLMs produced distinct interpretations and descriptions of genetic interactions. To combine the strength of multiple LLMs, we designed an ensemble structure that consists of three LLMs. In preliminary experiment, we found that large LLMs performed superiorly while acting as a generator. Therefore, the three LLM generators we used were GPT-4o, Gemini-2.0-Flash, and DeepSeek-V3 (Hurst et al., 2024; Liu et al., 2024; Team et al., 2024). Each LLM was independently prompted in a zero-shot setting. Each prompt consists of structured evidence from STRING and BioGRID databases, excerpts from PubMed abstracts, and an instruction to produce a structured object with the following major components, as shown in Figure 1:

- The strength of the interaction at three levels: strong, weak, or none.
- A brief description of the interaction.
- Any major diseases associated with this interaction.

3.3. LLM-as-a-Judge Evaluation

We found in our preliminary experiment that even human experts have difficulty to deterministically judge whether one description is more accurate and scientifically sound than others without consulting external resources, such as existing databases and related literature. Due to the large volume of data we generated, direct annotations by human experts are infeasible. To efficiently identify the best gene-gene interaction description, we employ a reasoning LLM model, referred to as LLM-as-a-Judge. Using LLMs as evaluation judges are common practices to improve text generation in domain knowledge (Gu et al., 2024). When at least one "non-none" interaction is generated by the LLM generators, the LLM judge evaluates the outputs from all three candidate models using a structured evaluation prompt. The judge assesses each description based on three criteria: factual accuracy, biological relevance, and writing clarity. The

LLM judge is instructed to return a JSON object indicating the selected candidate with a brief justification.

4. Experiment

We generated gene-gene interaction descriptions for 84,233 gene pairs across 2,000 genes, selected based on the highest variance in the TCGA Pan-cancer dataset via the Xena platform (Goldman et al., 2020). Among the generated interactions, 1,753 gene pairs have at least one non-null output from the LLM generators. After evaluation by LLM-as-a-judge, 506 interactions were identified with a strength level other than "none". The LLM-as-a-Judge was implemented using a zero-shot DeepSeek Distilled 32B model run locally on a server with two RTX 5090 GPUs (Guo et al., 2025).

4.1. LLM-GeneGraph Vs. Related Biology Knowledge Data

A demonstration of the interactions generated by LLM-GeneGraph and the corresponding BioGrid and STRING scores, is shown in Table 1. The comparison highlights the dramatic difference in the amount of information provided between traditional gene-gene interaction databases, like the BioGrid and STRING databases and the descriptive interaction information generated by the LLM-GeneGraph. By providing detailed textual context, LLM-GeneGraph serves as a foundation for future single-cell and spatial transcriptomics analysis tools built upon the generated textual knowledge graphs.

The percentiles of the BioGrid and STRING scores for all generated connections are summarized in Table 2. While more than half of the interactions have high-confidence STRING scores (> 0.7), more than 75% of the interactions have a BioGrid score of 0. This discrepancy showed the limitation of the current biological database, static nature and incompleteness.

Table 2. The Percentiles of BioGrid and STRING Scores for the Generated Interactions

Percentile	BioGrid	STRING
25th	0	0.429
50th	0	0.732
75th	0	0.843
90th	0.125	0.938
100th	0.314	0.999

4.2. Scientific Validation on New Found Interactions

To further access the scientific validity of the interactions identified by LLM-GeneGraph, we examined the interactions with both BioGrid and STRING scores equal to zero. There are nine such interactions, all classified as weak. For six out of nine interactions, we were able to find supporting evidence in the literature suggesting some level of association between the corresponding gene pairs, as shown in Table 3. These findings indicate that LLM-GeneGraph is capable of finding biologically plausible interactions that are omitted by existing databases, which concludes that LLM-GeneGraph outputs have relatively high scientific accuracy.

Table 3. Generated Interactions Not in BioGrid and STRING Databases

Gene 1	**Gene 2**	**Evidence**
C8A	CR2	N/A
IL8	WNT2	(Lin et al., 2022a; 2024)
DES	MUC4	(Forgó et al., 2021)
DES	UGT1A9	N/A
EMX2	EMX2OS	(Spigoni et al., 2010)
GC	REG4	(Rowe et al., 2020)
CP	FOXE1	(Moreno et al., 2009)
OLIG2	TF	(Cheli et al., 2023)
CCL21	PAH	N/A

Using the associated disease information, we identified 103 interactions linked to cancer. We computed the Pearson correlation coefficients for all 103 interactions using TCGA Pan-cancer expression data. The results showed an average correlation coefficient of 0.55 and a median of 0.583, indicating a moderate expression correlation in cancer-related data. Therefore, we can conclude that LLM-GeneGraph is not only able to generate genetic interactions consistent with current biological knowledge of genome, like BioGrid and STRING, but also new interactions that are beyond the scope of traditional databases as shown in Table 3.

5. Conclusion

In this work, we proposed LLM-GeneGraph, a zero-shot LLM-based framework designed to dynamically generate descriptive gene-gene interaction networks. It utilizes an ensemble structure composed of three state-of-the-art LLMs, and incorporates a reasoning LLM as a judge to select the best generated interaction description based on actual accuracy, biological relevance, and writing clarity.

We generated 84,233 candidate gene-gene interactions across 2,000 genes. After the evaluation of the LLM judge, the final data consists of 506 interactions with a strength level other than "none". We compared the generated interactions against BioGrid and STRING scores and concluded that most of the interactions are supported by the current gene-gene interaction databases. We identified nine new interactions that are not in BioGrid norm mn STRING databases. For six out of the nine interactions, we found supporting evidence, suggesting a degree of association between the gene pair. Therefore, we can conclude that LLM-GeneGraph can construct a scientifically accurate and biologically validated textual gene-gene interaction network, while discovering new insights of genomics beyond the scope of traditional datasets.

Software and Data

The generation code, all generated interactions, and final interactions filtered by the LLM-as-a-Judge are available at https://github.com/NabaviLab/LLM-GeneGraph.

Acknowledgements

This work is supported by the National Science Foundation (NSF) under grant No. 1942303, PI: Nabavi.

Impact Statement

This paper presents work whose goal is to advance the field of biomedical research and generative AI. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ayala, O. and Bechard, P. Reducing hallucination in structured outputs via retrieval-augmented generation. In Yang, Y., Davani, A., Sil, A., and Kumar, A. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 228–238, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-industry.19. URL https://aclanthology.org/2024.naacl-industry.19/.

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596, 2023.

Bertin, P., Hashir, M., Weiss, M., Frappier, V., Perkins, T. J., Boucher, G., and Cohen, J. P. Analysis of gene interaction graphs as prior knowledge for machine learning models. *arXiv preprint arXiv:1905.02960*, 2019.

Cheli, V. T., González, D. A. S., Wan, R., Rosenblum, S. L., Denaroso, G. E., Angeliu, C. G., Smith, Z., Wang, C., and Paez, P. M. Transferrin receptor is necessary for proper oligodendrocyte iron homeostasis and development. *Journal of Neuroscience*, 43(20):3614–3629, 2023.

Chen, Y. and Zou, J. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pp. 2023–10, 2024.

Feng, Y., Zhou, L., Ma, C., Zheng, Y., He, R., and Li, Y. Knowledge graph–based thought: a knowledge graph–enhanced llm framework for pan-cancer question answering. *GigaScience*, 2025. doi: 10.1093/gigascience/giae082.

Forgó, E., Hornick, J. L., and Charville, G. W. Muc4 is expressed in alveolar rhabdomyosarcoma. *Histopathology*, 78(6):905–908, 2021.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, 38(6):675–678, 2020.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

He, X., Tian, Y., Sun, Y., Chawla, N., Laurent, T., LeCun, Y., Bresson, X., and Hooi, B. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907, 2024.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Li, B. and Nabavi, S. scgemoc, a graph embedded contrastive learning single-cell multiomics clustering model. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2075–2080. IEEE, 2023.

Li, B. and Nabavi, S. A multimodal graph neural network framework for cancer molecular subtype classification. *BMC bioinformatics*, 25(1):27, 2024.

Li, Z., Gou, Z., Zhang, X., Liu, Z., Li, S., Hu, Y., Ling, C., Zhang, Z., and Zhao, L. Teg-db: A comprehensive dataset and benchmark of textual-edge graphs. *Advances in Neural Information Processing Systems*, 37:60980–60998, 2024.

Lin, L., Li, L., Ma, G., Kang, Y., Wang, X., and He, J. Overexpression of il-8 and wnt2 is associated with prognosis of gastric cancer. *Folia Histochemica et Cytobiologica*, 60(1):66–73, 2022a.

Lin, L., Tao, R., Cai, Y., Zhao, Q., Yang, K., Yang, W., and Guo, F. Expression and clinical significance of il-8 and wnt2 in helicobacter pylori-infected gastric cancer patients. *The Journal of Infection in Developing Countries*, 18(10):1512–1521, 2024.

Lin, X., Tian, T., Wei, Z., and Hakonarson, H. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature communications*, 13(1):7705, 2022b.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Moreno, L. M., Mansilla, M. A., Bullard, S. A., Cooper, M. E., Busch, T. D., Machida, J., Johnson, M. K., Brauer, D., Krahn, K., Daack-Hirsch, S., et al. Foxe1 association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. *Human molecular genetics*, 18 (24):4879–4896, 2009.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200, 2021.

Rosenbaum, G. R., Jiang, L. Y., Sheth, I., Stryker, J., Alyakin, A., Alber, D., Goff, N. K., Joon, Y., Kwon, H.-H., Markert, J., Nasir-Moin, M., Niehues, J., Sangwon, K. L., Yang, E., and Oermann, E. K. Medg-krp: Medical graph knowledge representation probing. *arXiv preprint arXiv:2412.10982*, 2024.

Rowe, M., Whittington, E., Borziak, K., Ravinet, M., Eroukhmanoff, F., Sætre, G.-P., and Dorus, S. Molecular diversification of the seminal fluid proteome in a recently diverged passerine species pair. *Molecular Biology and Evolution*, 37(2):488–506, 2020.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. Retrieval augmentation reduces hallucination in conversation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.320. URL https://aclanthology.org/2021.findings-emnlp.320/.

Spigoni, G., Gedressi, C., and Mallamaci, A. Regulation of emx2 expression by antisense transcripts in murine cortico-cerebral precursors. *PloS one*, 5(1):e8658, 2010.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Tianxiang Jin, Yingtong Dou, Haochen Chen, Suhang Wang, and Carl Yang. Edgeformers: Graph-empowered transformers on textual-edge networks. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 1–17, 2023. URL https://arxiv.org/abs/2302.11050.

Wang, K., Ding, Y., and Han, S. C. Graph neural networks for text classification: A survey. *Artificial Intelligence Review*, 57(8):190, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zhao, W., Gu, X., Chen, S., Wu, J., and Zhou, Z. Modig: integrating multi-omics and multi-dimensional gene network for cancer driver gene identification based on graph attention network model. *Bioinformatics*, 38(21):4901–4907, 2022.

Zhu, Y., Wang, Y., Shi, H., and Tang, S. Efficient tuning and inference for large language models on textual graphs. *arXiv preprint arXiv:2401.15569*, 2024.