

Automated Extraction of Molecular Interactions and Pathway Knowledge using Large Language Model, Galactica: Opportunities and Challenges

Gilchan Park*, Byung-Jun Yoon, Xihai Luo, Vanessa López-Marrero, Patrick Johnstone, Shinjae Yoo, Francis J. Alexander

Computational Science Initiative, Brookhaven National Laboratory, Upton, NY
{gpark, byoon, xluo, vlopezmar, pjohnston, sjyoo, falexander}@bnl.gov

Abstract

Understanding protein interactions and pathway knowledge is essential for comprehending living systems and investigating the mechanisms underlying various biological functions and complex diseases. While numerous databases curate such biological data obtained from literature and other sources, they are not comprehensive and require considerable effort to maintain. One mitigation strategies can be utilizing large language models to automatically extract biological information and explore their potential in life science research. This study presents an initial investigation of the efficacy of utilizing a large language model, Galactica in life science research by assessing its performance on tasks involving protein interactions, pathways, and gene regulatory relation recognition. The paper details the results obtained from the model evaluation, highlights the findings, and discusses the opportunities and challenges. The code and data are available at: <https://github.com/boxorange/BioIE-LLM>

1 Introduction

A significant portion of contemporary molecular biology research is dedicated to studying and comprehending the roles and interactions of the countless proteins that form the fundamental building blocks of life. The prediction of protein structures and functions is essential in addressing crucial challenges in life science, including developing therapeutic solutions for various diseases. By speeding up drug discovery and development, such advancements could significantly enhance healthcare. The majority of proteins have undefined functions, and only a fraction of them have been unequivocally identified through arduous and intensive laboratory research. These established protein functions

are used as a benchmark to predict functions computationally by analyzing DNA and amino acid sequence homology across the continuously growing repository of protein sequences obtained from genome sequencing. To deeply understand protein functions, protein interaction information can be crucial, and many databases such as STRING, KEGG, IntAct, BioGrid, DIP, and HPRD have been established to gather and maintain pathway analysis and regulatory results obtained by lab experiments and from the scientific literature. Regrettably, extracting information from the existing literature demands significant manual labor and is a time-consuming process. One viable solution to remedy this is to leverage efficient machine learning models that can accurately recognize such information in scientific texts.

In recent years, large language models (LLMs) have gained significant attention in the natural language processing (NLP) field owing to their capability to execute complex language tasks, their flexibility, and their potential to generate responses similar to humans (Brown et al., 2020; Zhao et al., 2023). Their application in various domains and tasks, including knowledge extraction from texts, has yielded promising outcomes (Agrawal et al., 2022). Our study aims to investigate the potential of LLMs in extracting pathway knowledge, protein interaction, and gene regulatory information. In this study, we have assessed the capability of Galactica (Taylor et al., 2022), a general-purpose scientific LLM, to accomplish these biological tasks. Although Galactica did not yield optimal outcomes in our biology-related tasks, it exhibited the capacity to identify specific genes/proteins, pathways, and their interactions. Our preliminary findings regarding this evaluation are presented in this paper.

2 Related Work

The field of biology encompasses challenging tasks such as identifying protein-protein interactions

*Corresponding author.

(PPIs) and gene regulatory relations. Additionally, pathway analysis is a crucial area of study as it documents the interactions between proteins and reflects important molecular biological processes, such as metabolic, signaling, protein interaction, and gene regulation processes. Research such as that for expression-based disease diagnosis (Lee et al., 2008; Gatza et al., 2010) and the identification of disease markers (Khunlertgit and Yoon, 2016) suggests that tasks based on pathway activity can be more stable than tasks based solely on genes. The scientific literature of biological sciences serves as an important repository of essential knowledge that has yet to be effectively discovered. To address this, NLP models based on deep neural networks have been widely adopted for analysis of structural properties of proteins (Vig et al., 2020), PPIs (Peng and Lu, 2017; Park et al., 2022), and pathway analysis (Casani-Galdón et al., 2020).

Several studies showed that LLMs performed comparably to traditional neural network models that necessitate labeled training data and fine-tuning processes, resulting in significant time and effort savings while providing a universal model capable of managing multiple tasks simultaneously (Kojima et al., 2022; Yuan and Liu, 2022). The Galactica LLM (Taylor et al., 2022) has been trained on a massive amount of scientific literature and has successfully tackled biological understanding task such as sequence validation perplexity, functional keyword prediction, protein function description. Hence, we aim to further examine the potential of Galactica in the domain of biological scientific knowledge.

3 Experiments

We investigated the potential of Galactica for addressing biological tasks related to PPIs, pathway knowledge, and gene regulatory relations. To accomplish this, STRING, KEGG, INDRA databases were adopted, and the details pertaining to the utilized data for these tasks can be found in Appendix A. In the context of a LLM, the proper selection of the number of examples or shots is essential to ensure efficient engineering. For this purpose, an ablation study was conducted to identify the optimal number of shots for each task. The shot number associated with the highest performance in test samples was selected for implementation, as detailed in Appendix B. Additionally, prompt construction is another critical factor that merits

	1K	10K
Galactica (6.7B)	0.166	0.161

Table 1: STRING Task 1 - Precision for the generated binding proteins for 1K/10K protein samples.

	1K	10K	100K
Galactica (6.7B)	0.552	0.558	0.562

Table 2: STRING Task 2 - Micro F-scores for randomly selected 1K/10K/100K positive and negative protein pairs (I.e., 1K = 500 pos + 500 neg).

attention, and the prompts tested for each task are listed in Appendix C. The experimental setup is detailed in Appendix D.

3.1 Recognizing Protein-Protein Interactions

We evaluated the Galactica on protein binding information recognition on a human protein network from the STRING DB. Specifically, we employed the model to produce a list of proteins that bind to a designated protein, as part of the generative question task (*STRING Task 1: generative question*).

<Predicted answer by model>
 Question: Which proteins are related to TBC1D9?
 Answer: TBC1D8, TBC1D14, TBC1D7, TBC1D5, TBC1D6, TBC1D

<Actual answer>
 Answer: TBC1D8, TBC1D14, TBC1D7, TBC1D5, PLK5, MYO16

To assess performance, we randomly selected 1,000 and 10,000 samples from the network for testing. The generated binding proteins were matched with the proteins in the network with an approximately 0.16 precision as described in Table 1. The results of the prediction analysis indicated that the model exhibited a tendency to generate words primarily from the initial letters of a given protein. Consequently, the accuracy of the predictions was considerably high for proteins with similar names, such as IKZF4 and RFC5, while a significant mismatch between predicted and actual binding proteins was observed in cases where dissimilar protein names were involved, such as DNAJC10 and TRIP11. The details of those examples are provided in Appendix E.

Subsequently, we tested the model’s recognition of protein binding relationships in a binary setting, which was formulated as a yes/no question to determine if two proteins bind to each other (*STRING Task 2: yes/no question*).

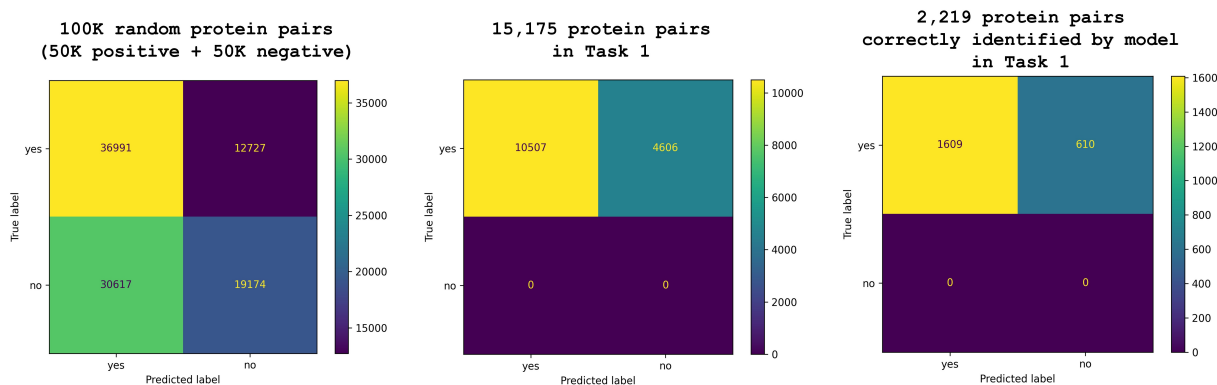


Figure 1: Confusion matrices for *STRING Task 2* (The observed disparity between the total number of samples and the sum of values in the confusion matrix can be attributed to the omission of responses other than 'yes' or 'no'.)

	Task1 pairs [†]	Consistency [‡]
Galactica (6.7B)	0.691	0.726

Table 3: *STRING Task 2* - Precision for the protein pairs used in *STRING Task 1*. [†]All positive protein pairs. [‡]Model prediction consistency between Task1 and Task2.

<Predicted answer by model>
 Question: Are CHEK2 and BRCA2 related to each other?
 Answer: **yes**

<Actual answer>
 Answer: **yes**

To create negative protein binding pairs, we utilized unconnected pairs from the human protein network. The experiment was conducted on 1K/10K/100K protein pairs, and the model exhibited F-scores slightly exceeding 0.5 in its predictions as described in Table 2. In order to evaluate the consistency of the model’s predictions across Task1 and Task2, we conducted an assessment of *STRING Task 2* using the identical protein pairs employed in *STRING Task 1*. These pairs exclusively consisted of positive instances. More specifically, our examination focused on whether the model successfully generated protein A associated with protein B and correctly classified their relationship as 'yes'. To this end, we tested *STRING Task 2* using the 1K protein pairs used in *STRING Task 1*, and the model performed *STRING Task 2* on all positive protein pairs and those generated correctly by the models in *STRING Task 1*. The evaluation results and the confusion matrices are described in Table 3 and Figure 1. The model exhibited variability in its predictive performance between *STRING Task 1* and *STRING Task 2* (0.73).

Although this precision (0.73) is slightly higher than the precision of all positive protein pairs used in *STRING Task 1*, this level of precision aligns closely with the precision observed for 50K random positive pairs (0.74 = 36,991 true positives / 50,000 true positive + false positives). This implies that while the model successfully made some accurate predictions for proteins in generative questions, it encountered challenges when identifying protein relations in the binary format prompts.

3.2 KEGG Pathway Recognition

Galactica was assessed for its ability to recognize genes related to human pathways associated with low-dose radiation exposure in KEGG. The task involved generating a list of genes belonging to the top 20 human pathways linked to low-dose radiation exposure (*KEGG Task 1: generative question*).

<Predicted answer by model>
 Question: Which genes are involved in "Adherens junction"??
 Answer: **CDH1, CTNNA3, CTNNB1, CTNNA1, CTNNA2, CTNNA8, CTNNA15**

<Actual answer>
 Answer: **CDH1, CTNNA3, CTNNB1, CTNNA1, CTNNA2, TGF1a, MEKK7**

The model accurately predicted the genes that belong to the pathways with about precision 0.26 as presented in Table 4, which outperformed the previous *STRING Task 1* generative test. Our speculation is that the model’s superior performance in recognizing low-dose radiation-related pathways compared to proteins might be attributed to the fact that pathway names related to low-dose radiation exposure are typically mentioned in specific sections or categories, whereas protein names are

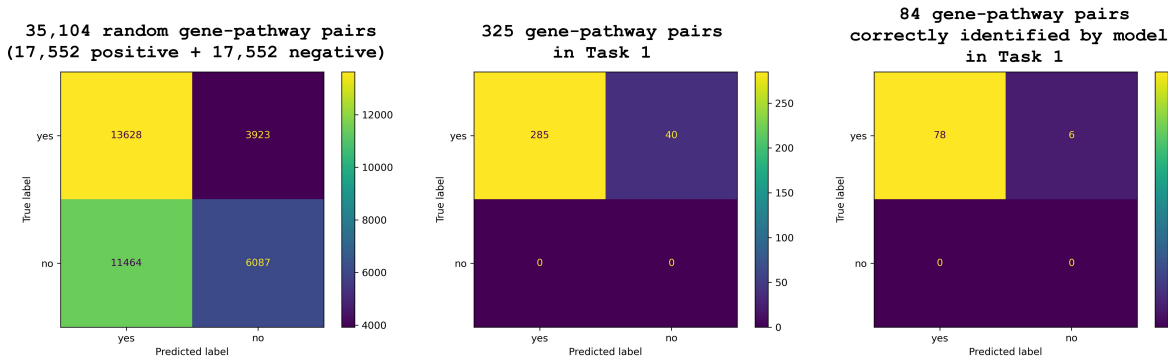


Figure 2: Confusion matrices for *KEGG Task 2*.

Pathways	
Galactica (6.7B)	0.256

Table 4: KEGG Task 1 - Precision for the generated genes that belong to the top 20 pathways relevant to low-dose radiation exposure.

35,174 gene and pathway pairs	
Galactica (6.7B)	0.562

Table 5: KEGG Task 2 - Micro F-scores for all positive and negative pairs (35,174 = 17,587 pos + 17,587 neg).

	Task1 pairs [†]	Consistency [‡]
Galactica (6.7B)	0.883	0.917

Table 6: KEGG Task 2 - Precision for the gene-pathway pairs used in KEGG Task 1. [†]All positive gene-pathway pairs. [‡]Model prediction consistency between Task1 and Task2.

more commonly found in a wider range of topic papers. This suggests that searching for information in a well-defined collection of data may yield more precise results than searching for information derived from ambiguous inputs in heterogeneous sources. The outcome of the prediction analysis showed that the genes produced for a particular pathway exhibited comparable patterns, a finding which had also been observed in the prior *STRING Task 1* experiment. Examples of this can be found in Appendix F.

We performed yes/no questions for pathways and genes relation recognition (*KEGG Task 2: yes/no question*). Similar to the *STRING Task 2*, we used member genes in other pathways as negative samples for a given pathway if they do not appear in the pathway. The model was evaluated on all positive relations (+ randomly chosen negative relations) and the relations used in *KEGG Task 1*

to measure the model consistency between Task1 and Task2.

<Predicted answer by model>
 Question: Are "DP beta 1" and "Type I diabetes mellitus" related to each other?
 Answer: yes

<Actual answer>
 Answer: yes

The results and the confusion matrices are displayed in Table 5, Table 6, and Figure 2 respectively. The model achieved an F-score of approximately 0.56 when making predictions for both positive and negative gene-pathway pairs. The precision of 0.92 for the model consistency signifies the model's reliable predictive performance. Moreover, the higher score achieved for the generated positive relationships (0.92) in contrast to the scores for all positive relationships in the pathways (0.78 = 13,628 / 17,552) and Task 1 pairs (0.88) suggests that the model possesses a greater level of comprehension concerning specific pathways compared to others.

In the *STRING Task 2* and *KEGG Task 2*, the model's responses to yes/no questions utilizing positive and negative samples skewed more towards positive, as illustrated by the leftmost confusion matrix in Figures 1 and 2. A plausible explanation for this outcome is the likelihood of erroneous negative relationships in the negative samples. For instance, among the negative samples is the relationship between the gene "HD1" and the pathway "Adherens junction" despite the fact that they are genuinely connected.

3.3 Evaluating Gene Regulatory Relations

Finally, we examined Galactica's ability to recognize human gene regulatory relations using data from the INDRA DB. Unlike the previous datasets,

	2	3	4	5	6
Galactica 6.7B	0.704	0.605	0.567	0.585	0.597

Table 7: INDRA Task - Micro F-scores with 1K samples for each class. See Appendix G for class names.

INDRA statements not only provide relation entities but also text snippets from research papers. We used these text snippets as contextual information about regulatory relations to generate questions for the model. The task involved asking the model to select the correct relationship between two genes from multiple relation classes in a given text (*INDRA Task: multiple-choice question*). This task serves as an evaluation of the model’s reading comprehension skills specifically related to gene regulatory relation texts.

<Predicted answer by model>
 Upon binding with Shh, Ptc1 inactivation allows Smo to initiate signaling XREF_BIBR, XREF_BIBR, XREF_BIBR through the Gli family of transcription factors.

Question: Given the options: "Activation", "Inhibition", "Phosphorylation", "Dephosphorylation", "Ubiquitination", "Deubiquitination", which one is the relation type between Ptc1 and Smo in the text above?

Answer: [Activation](#)

<Actual answer>
 Answer: [Activation](#)

To construct multiple-choice questions, we identified the six most frequently occurring classes in the dataset and utilized two to six of them for choices. The model was assessed using 1K samples for each class, and the findings are detailed in Table 7 and Figure 3. With the escalation of problem complexity due to the increased number of choices, the model initially encountered difficulties in identifying correct answers. Nonetheless, it exhibited improved performance in six-choice problems compared to cases involving four or five choices. When examining the results of yes/no questions in *STRING Task 2* and *KEGG Task 2* (≈ 0.56) using the two-class F-score (0.70) in *INDRA Task*, it becomes evident that the model possesses a more consistent ability to recognize entity relations within contexts as compared to extracting information through straightforward questioning. This observation suggests that incorporating contextual information in questions could potentially enhance the model’s predictive capabilities.

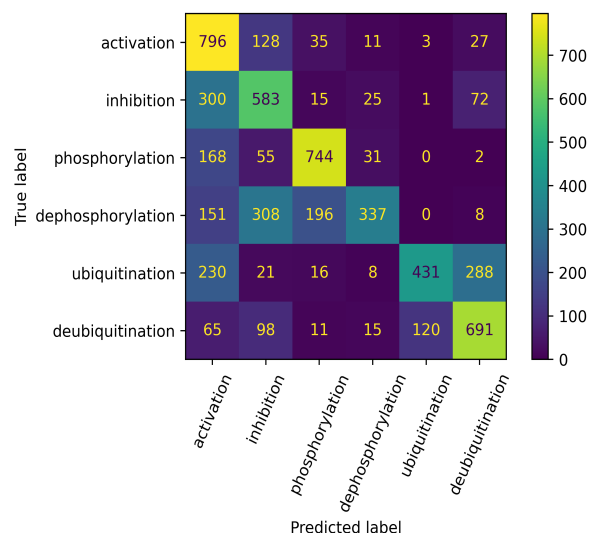


Figure 3: Confusion matrix of 6 choice questions for INDRA Task.

4 Discussion and Conclusion

This study evaluated the performance of Galactica on various biological tasks using different types of database resources, including PPIs, pathway knowledge, and regulatory relations. Based on our findings, it was observed that the model faced greater difficulty in answering questions that contained limited information, in contrast to questions that provided contextual details and were more specific. While the Galactica did not produce ideal results in our biology-related tasks, we observed that the model demonstrated the ability to recognize certain genes and proteins and their interactions. Despite this, our approach presents a potential avenue for using the model, and we anticipate that our findings will assist domain scientists and researchers in employing the model for their applications and obtaining insights into the model’s behavior based on experimental outcomes.

5 Work in progress

Our study suggests that the model can show better performances when contextual text is provided. We will consider forging a chain of queries to generate context for the datasets not having supporting text (See Appendix H). We found that the model’s performance was largely affected by prompts. This needs to be further investigated. We plan to evaluate other LLMs such as GPT series, LLaMA, and Alpaca by comparing with smaller sized domain-specifically trained models such as BioGPT and BioMedLM in biological tasks.

Limitations

There are a few limitations for our work. First, to assess the model’s ability to recognize negative protein-protein and gene-pathway pairs, we used unconnected pairs in the datasets. However, the negative pairs have not been proved, which might contain authentic or potential positive connections. We continuously search for truly negative gene/protein pairs established by experiments and research. Second, this work only reports the assessment of the Galactica standard model (6.7B). The evaluation of the other models (mini: 125 M, base: 1.3 B, large: 30 B, huge: 120 B) remains as future works. In-depth studies on the model inference will be also followed in our future works such as clustering of PPIs and genes belonging to pathways predicted by the model.

Acknowledgements

This work is supported by the U.S. Department of Energy, Office of Science, RadBio program under Award KP1601011/KP1601017/FWP CC121.

References

- Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Salvador Casanı-Galdon, Cecile Pereira, and Ana Conesa. 2020. Padhoc: a computational pipeline for pathway reconstruction on the fly. *Bioinformatics*, 36(Supplement_2):i795–i803.
- Michael L Gatz, Joseph E Lucas, William T Barry, Jong Wook Kim, Quanli Wang, Matthew D Crawford, Michael B Datto, Michael Kelley, Bernard Mathey-Prevot, Anil Potti, et al. 2010. A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences*, 107(15):6994–6999.
- Benjamin M Gyori, John A Bachman, Kartik Subramanian, Jeremy L Muhlich, Lucian Galescu, and Peter K Sorger. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular systems biology*, 13(11):954.
- Minoru Kanehisa and Susumu Goto. 2000. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.
- Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. 2019. New approach for understanding genome variations in kegg. *Nucleic acids research*, 47(D1):D590–D595.
- Navadon Khunlertgit and Byung-Jun Yoon. 2016. Incorporating topological information for predicting robust cancer subnetwork markers in human protein-protein interaction network. In *BMC bioinformatics*, volume 17, pages 143–152. Springer.
- Hyunwoong Ko. 2021. Parallelfomers: An efficient model parallelization toolkit for deployment. <https://github.com/tunib-ai/parallelfomers>.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.
- Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. 2008. Inferring pathway activity toward precise disease classification. *PLoS computational biology*, 4(11):e1000217.
- Xihaier Luo, Sean McCorkle, Gilchan Park, Vanessa Lopez-Marrero, Shinjae Yoo, Edward R Dougherty, Xiaoning Qian, Francis J Alexander, and Byung-Jun Yoon. 2022. Comprehensive analysis of gene expression profiles to radiation exposure reveals molecular signatures of low-dose radiation response. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2366–2374. IEEE.
- Gilchan Park, Sean McCorkle, Carlos Soto, Ian Blaby, and Shinjae Yoo. 2022. Extracting protein-protein interactions (ppis) from biomedical literature using attention-based relational context information. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2052–2061. IEEE.
- Yifan Peng and Zhiyong Lu. 2017. Deep learning for extracting protein-protein interactions from biomedical literature. *BioNLP 2017*, page 29.
- Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. 2021. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612.
- 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 preprint arXiv:2211.09085*.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Nazneen Rajani, et al. 2020. Bertology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Weizhe Yuan and Pengfei Liu. 2022. restructured pre-training. *arXiv preprint arXiv:2206.11147*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Data Description

STRING DB (Szklarczyk et al., 2021): The present study employed the human (*Homo sapiens*) protein network for performing a protein-protein interaction (PPI) recognition task. The network was constructed based on the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database, which is a comprehensive biological repository and online resource for both predicted and confirmed protein interactions. The database integrates data from a range of sources, including experimental studies, computational prediction methods, and publicly available text collections. The human network encompasses 19,566 proteins and 5,968,680 protein bindings.

KEGG DB (Kanehisa and Goto, 2000): The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a set of databases encompassing a wide range of biological information, including genomic data, disease information, chemical compounds, and biological pathways. It houses a staggering collection of over 28,000 complete genomes, encompassing a diverse range of organisms. Furthermore, it hosts an expansive repertoire of more than 500 pathways, meticulously curated and annotated to illuminate the intricate web of molecular interactions that govern various biological processes. Moreover, the database includes approximately 5 million reference genes, providing researchers with invaluable resources

for gene-centric investigations (Kanehisa et al., 2019). The KEGG pathways contain molecular interactions and reactions, which are designed to link genes in the genome to gene products (mostly proteins) in biological pathways. The focus of our investigation pertains to the pathways within the human body that are affected by exposure to low-dose ionizing radiation, which remains a significant threat to human health and is not yet fully comprehended. To explore this topic, we utilized the KEGG human pathways which have been identified as being activated in response to low-dose radiation exposure in a recent study (Luo et al., 2022).

INDRA DB (Gyori et al., 2017): The Integrated Network and Dynamical Reasoning Assembler (INDRA) is a tool that facilitates the integration of information regarding causal mechanisms into a unified format suitable for the construction of a variety of predictive and explanatory models. In the field of molecular biology, sources of mechanistic information include pathway databases, textual descriptions of mechanisms generated by human curators, and information extracted from the scientific literature through text mining. The INDRA platform streamlines this information by removing duplicates, standardizing the data, and organizing it into a set of Statements accompanied by associated evidence. By collating information from multiple sources in this manner, INDRA enables researchers to build robust models for exploring the complex molecular mechanisms underlying biological systems. The present study utilized a set of human gene regulatory relation statements that represent mechanistic interactions between biological agents. The dataset comprises a total of 4,258,718 distinct statements.

B Ablation study on the number of shots

1. STRING Task 1: We randomly drew 1K and 10K samples out of the STRING DB human protein network for testing, and the generated binding proteins corresponded to the proteins in the human network the most with 1-shot prompting as seen in Table 8
2. STRING Task 2: We evaluated 1K samples (500 true cases + 500 false cases) randomly drawn from the STRING DB human protein network with different number of prompt

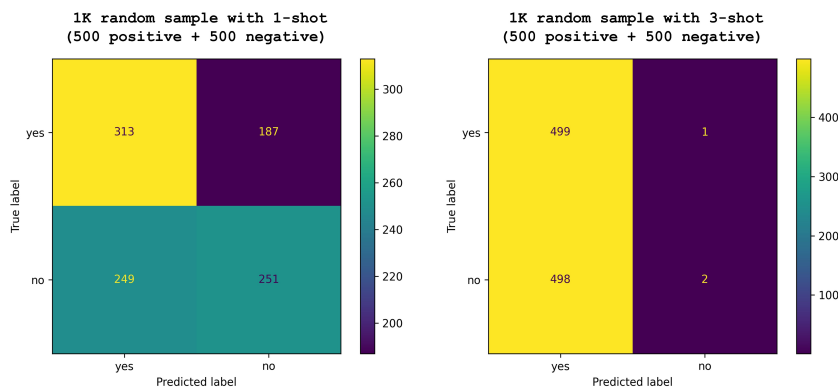


Figure 4: Confusion matrices for KEGG Pathway Recognition Task 2 ablation study.

		0-shot	1-shot	2-shot	3-shot
Precision	1K	0.127	0.166	0.145	0.135
	10K	0.130	0.144	0.140	0.137

Table 8: Precision of different shots with 1K/10K samples for STRING Task 1 using a human protein network from STRING DB.

shots. Here, N -shot indicates the combination of N number of true samples and N number of false samples (e.g., 1-shot: 1 true + 1 false (total 2 samples)). The result showed that 3-shot prompt performed the best in Table 9.

- KEGG Pathway Recognition Task 1: We assessed human pathways associated with low-dose radiation exposure in KEGG DB with different number of shots, and 1-shot prompting showed the best performance as described in Table 10.
- KEGG Pathway Recognition Task 2: We evaluated 1K samples (500 true cases + 500 false cases) randomly drawn from human pathways associated with low-dose radiation exposure in KEGG DB with different number of prompt shots. Here, N -shot indicates the combination of N number of true samples and N number of false samples (e.g., 1-shot: 1 true + 1 false (total 2 samples)). The result showed that 1-shot prompt performed the best in Table 11.
- Evaluating Gene Regulatory Relations Task: We tested different shots with 400 samples for 4 classes (100 Activation + 100 Inhibition + 100 Phosphorylation + 100 Dephosphorylation) from INDRA DB, and 2-shot prompting showed the best performance on the multiple choice task in Table 12.

	0-shot	1-shot	2-shot	3-shot
Micro F1	0.515	0.552	0.543	0.590 [†]

Table 9: Micro F-scores of different shots with 1K samples for STRING Task 2 using a human protein network from STRING DB. [†] Due to the high false positive rate, 1-shot prompting was adopted.

	0-shot	1-shot	2-shot	3-shot
Precision	0.170	0.259	0.221	0.209

Table 10: Precision of different shots for KEGG Pathway Recognition Task 1.

	0-shot	1-shot	2-shot	3-shot
Micro F1	0.489	0.564	0.534	0.501

Table 11: Micro F-scores of different shots with 1K samples for KEGG Pathway Recognition Task 2.

	0-shot	1-shot	2-shot	3-shot
Micro F1	0.370	0.508	0.610	0.560

Table 12: Micro F-scores of different shots with 400 samples (100 Activation + 100 Inhibition + 100 Phosphorylation + 100 Dephosphorylation) choice for Evaluating Gene Regulatory Relations Task using INDRA DB.

C Tested Prompts

STRING Task 1:

1. "Which proteins are bound to x?"
2. "What proteins are bound to x?"
3. "What proteins are bound to x?"
4. "What proteins does x bind to?"
5. "To what proteins does x bind?"
6. "Which proteins are related to x?"

STRING Task 2:

1. "Do the two proteins "x" and "y" bind each other?"
2. "Do the two proteins x and y bind each other? True or False"
3. "Do the two proteins x and y bind to each other? True or False"
4. "Do x and y bind each other? True or False"
5. "Does x bind to y? True or False"
6. "Do x and y bind to each other? True or False"
7. "Are x and y related to each other? True or False"
8. "Are x and y related to each other?"
9. "Are x and y related to each other? yes or no"
10. "x and y are related to each other. Is this statement True or False?"
11. "x and y are related to each other."
12. "Given the options: "Related", "Unrelated", which one is the relation type between x and y?"

KEGG Pathway Recognition Task 1:

1. "Which genes are involved in "x"?"
2. "Which genes are involved in x?"
3. "Which genes are related to x?"
4. "Which proteins are related to x?"
5. "Which genes or proteins are related to x?"
6. "Which genes/proteins are related to x?"

KEGG Pathway Recognition Task 2:

1. "Are x and y related to each other?"
2. "Are "x" and "y" related to each other?"
3. "Is x related to y?"

4. "Is x related to the pathway y?"
5. "Is x involved in y?"
6. "Is "x" involved in "y"?"
7. "Is x involved in the human pathway y?"
8. "Is x involved in the KEGG pathway y?"
9. "Does "y" have "x"?"

Evaluating Gene Regulatory Relations Task:

1. "Which of the following is the relation type between x and y in the text above?"
2. "Which of the following is the relation type between "x" and "y" in the text above?"
3. "Which of the following is the relation between "x" and "y" in the text above?"
4. "Given the options: z, which one is the relation type between "x" and "y" in the text above?"
5. "Given the options: z, which one is the relation type between x and y in the text above?"

D Experimental Setup

We used the Galactica standard model with 6.7 billion parameters. The experiments were conducted on 8×NVIDIA V100 GPUs, and we exploited Galactica’s option for model tensor parallelism based on Parallelfomers (Ko, 2021) when the machine has enough memories, which significantly increases task processing time (about twice faster). The model processed a batch sized input for a task, which is the number of prompts to infer (I.e., the number of input texts for model generation at once). The batch sizes for the tasks are as follows.

- STRING Task 1 (generative question): 16, 32
- STRING Task 2 (yes/no question): 32, 64
- KEGG Task 1 (generative question): 16, 32
- KEGG Task 2 (yes/no question): 32, 64
- INDRA Task (multiple choice question): 4, 8

E STRING Task 1 examples

Question: Which proteins are related to IKZF4?
Answer (**pred**): IKZF1, IKZF2, IKZF3, IKZF5
Answer (**true**): IKZF1, IKZF2, IKZF3, IKZF5

Question: Which proteins are related to RFC5?
 Answer (**pred**): RFC1, RFC2, RFC3, RFC4, RFC5
 Answer (**true**): RFC1, RFC2, RFC3, RFC4, ERCC1

Question: Which proteins are related to USP32?
 Answer (**pred**): USP54, USP41, USP42, USP34, USP38, USP50, USP52, USP32, USP55, USP56, ...
 Answer (**true**): USP54, USP41, USP42, USP34, USP38, USP50, CACNA1H, ACTC1, DHX32, MAGI3, ...

Question: Which proteins are related to DNAJC10?
 Answer (**pred**): DNAJC10K, DNAJC100, DNAJC10D, DNAJC10Q, DNAJC10E ...
 Answer (**true**): ARL5B, MATN3, KPNA4, SRSF10, MTHFD2, ...

Question: Which proteins are related to TRIP11?
 Answer (**pred**): TRIP32, TRIP31, TRIP20, TRIP23, TRIP25, ...
 Answer (**true**): RIC1, GOLGA8J, ARL5B, TMTC4, KIF5A, ...

F KEGG Pathway Recognition Task 1 examples

Question: Which genes are involved in “Cellular senescence”?
 Answer (**pred**): CDKN2A, CDKN2B, CDKN1A, CDKN1D, CDKN1E, CDKN1B, CDKN1C
 Answer (**true**): CDKN2A, CDKN2B, CDKN1A, p107, major histocompatibility complex, mitogen-activated protein kinase 14, P44MAPK

Question: Which genes are involved in “Phagosome”?
 Answer (**pred**): RAB5A, RAB7A, RAB47, RAB41, RAB48, RAB39B, RAB50 ...
 Answer (**true**): RAB5A, RAB7A, ATP6G, CD51, HEL-S-62p, phospholipase A2 receptor 1, ATP6EL2, ...

Question: Which genes are involved in “Proteoglycans in cancer”?
 Answer (**pred**): CD63, CD284, CD282, CD44, CD166, CD276, CD278, CD81, CD55, ...
 Answer (**true**): CD63, CD284, CD282, CD44, SJS1, G17P1, GAB1, PLCE1, HPSE1, ...

Question: Which genes are involved in “Autoimmune thyroid disease”?
 Answer (**pred**): TSHR, TSH
 Answer (**true**): TSHR, hTSHR-I

G INDRA DB class names

Table 13 displays the name of classes in INDRA DB statements used in the Evaluating Gene Regu-

latory Relations Task.

# Choices	Classes
2 class	Activation, Inhibition
3 class	Activation, Inhibition, Phosphorylation
4 class	Activation, Inhibition, Phosphorylation, Dephosphorylation,
5 class	Activation, Inhibition, Phosphorylation, Dephosphorylation, Ubiquitination,
6 class	Activation, Inhibition, Phosphorylation, Dephosphorylation, Ubiquitination, Deubiquitination

Table 13: The class names used in the multiple choice question for Evaluating Gene Regulatory Relations Task using INDRA DB.

H A chain of query example

To provide a model with context information about a query, we plan to apply an iterative prompting for a chain of thought (CoT) development (Wei et al., 2022; Wang et al., 2022). An example is illustrated below.

Question: what is “Natural killer cell mediated cytotoxicity”?
 -> The answer of this query can be a context of the following query.

Question: Which genes are involved in “Natural killer cell mediated cytotoxicity”?
 Answer: VAV3, NFAT5, HCST, CHP1, SH2D1B, RAET1E