Jack and the BeansTALK: Towards Question Answering in Plant Biology

Gaganpreet Jhajj[1,3] and **Yumeko Nomura**[2,3]
[1]Athabasca University
[2]University of North Carolina at Chapel Hill
[3]Okinawa Institute of Science and Technology
gjhajj1@learn.athabascau.ca, nomuray@unc.edu

Abstract

Developing effective question-answering systems tailored to specific domains remains an active area of research in NLP. This paper details the first steps in creating BeansTALK, an innovative question-answering system explicitly designed for plant biology. LLMs have contributed to a rapid rise in consumer adoption of generative AI technologies; however, in specific research fields, their applications are still minimal. To address this, we propose BeansTALK, which utilizes a corpus of published articles to construct a knowledge graph via an LLM. This graph will form the core of our system, which incorporates Retrieval-Augmented Generation (RAG) to enhance answer precision and relevance. Future work will leverage BeansTALK's potential to assist with literature reviews, enhance accessibility in plant biology research, and stimulate collaborative endeavors within the discipline.

Keywords: Plant Biology, Knowledge Graphs, Question Answering

1 Introduction and Related Work

Developing robust question-answer (QA) systems is an ongoing area of research (Ojokoh and Adebisi, 2019). In NLP, new models and approaches to modeling domain knowledge and generative AI such as GPT (OpenAI, 2023) have made it easier to create structured knowledge representations for downstream tasks such as QA.

In biology, prior work on QA has been explored (Neves and Leser, 2015). However, there remains space within the sub-domain of plant biology where little work has been done on QA systems. While AI and machine learning techniques have become increasingly leveraged within plant sciences (Brink, 2024), there is currently a lack of work on knowledge domain modeling in plant biology.

Plant biology is a broad discipline that investigates traits from the molecular level to the ecosystem. Virtual assistants such as ChatGPT are convenient for searching for potential ideas or problem-solving. Various proposed systems currently leverage LLMs to create questions (Kabir and Lin, 2023; Rodrigues et al., 2024). However, such systems can fall prey to having gaps in knowledge that are needed for effective domain-specific applications. Given the knowledge gap for QA systems that exist in plant biology, we propose BeansTALK, a robust plant biology QA system, and detail the first step toward its implementation.

2 Methods

For this initial step towards implementing our program, we created a mini corpus of publications selected from Nature Plants[1]. We selected 99 open-access articles published between June 2016 and October 2024 as shown in Appendix A. Figure 1 showcases a preliminary overview of our process.

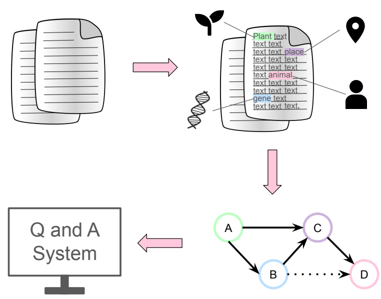


Figure 1: Initial steps in the creation of the QA system, BeansTALK. First, a corpus of open access papers was created. Then, we carried out entity relationship extraction via an LLM. Using these entity relations, we create a knowledge graph (KG) in Neo4j. This KG will be the basis for our future QA system.

After creating this mini corpus, we used GPT-4, GPT-4-Turbo, GPT-4o, and GPT-4o mini (OpenAI, 2023) via the OpenAI API with LangChain [2] to

[1]https://www.nature.com/nplants/
[2]https://www.langchain.com/

extract nodes and relationships for our KG, which will be the backend for our future QA system. Prior work has pointed to how LLMs are effective in constructing KGs (Jhajj et al., 2024; Trajanoska et al., 2023). Ongoing research continues to explore the effectiveness of triplet extraction via LLMs (Papaluca et al., 2024).

The KG, derived from the mini corpus, will form the foundation of our QA system. This graph will not only be a repository of information but also enable us to perform Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Siriwardhana et al., 2023), enhancing the system's capabilities.

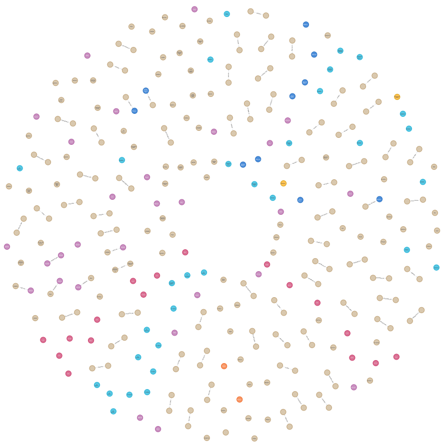


Figure 2: KG generated from the mini-corpus with Gpt-4o. All graphs can be seen in Appendix B.

	GPT-4	GPT-4 Turbo	GPT-4o	GPT-4o mini
Nodes	101	1614	2368	215
Relationships	28	30854	665	67
Labels	8	98	122	15
Relationship Types	11	151	226	30

Table 1: Comparison of KG characteristics for each model.

Using LLMs for entity extraction on our mini-corpus, we generated 101 nodes and 28 relationships using GPT-4 as seen in Table 1. Using Gpt-4o, we generated 2368 nodes, and part of the KG can be seen in Figure 2. Gpt-4 Turbo had the most relationships between nodes. However, this can be attributed to some error in the cypher generation process that created edges between many nodes in the graph structure through an ongoing loop. An example summary of the LLM suggested nodes from processed papers can be seen in Appendix C for one of the 99 papers.

3 Future Work

We created a small corpus of publications as a preliminary step in this work. We hope to add papers across multiple journals to develop a QA system that will inspire plant biologists and students in the field with accurate and concise concepts and ideas. We anticipate this will reduce literature review time and make the field more accessible to all, which can lead to increased collaboration and generate innovative ideas.

As seen in Figure 2 and Appendix B, our approach created a series of disjoint and distributed nodes in our KGs. Going forward, we aim to mediate this by leveraging a form of NER similar to BioNER (Perera et al., 2020) to improve entity node extraction and expand the corpus to include more academic articles and multimodal data. QA in science remains a distinctly different problem than QA more broadly, so we aim to benchmark our work against metrics such as SciQA (Auer et al., 2023). In addition, we plan to test our work on domain-relevant information extraction datasets such as the BioNLP Shared Task 2016 on Plant Seed Development (Chaix et al., 2016).

This work is still the groundwork and is in progress, so we aim to build an extensive KG that represents the broader field of plant biology and leverage our knowledge graph as the basis for this system (Huang et al., 2019). We will need to revisit this work and investigate more prompting approaches, as we currently prompted the models to extract "entities (e.g., species, genes, molecules) and relationships (e.g., interactions, regulations)," however, we chose to limit the scope for this initial exploration. Additionally, using tuned models such as in (Zhang et al., 2024) would most likely increase the effectiveness of our efforts.

4 Conclusion

QA systems and conversational AI tools are becoming increasingly common and ongoing research areas. This was the first step toward creating BeansTALK, a QA tool. We anticipate that BeansTALK will be a valuable tool for users to learn about plant biology and provide question-answering capabilities to researchers and students alike.

We are aware of the current limitations, but we expect to expand our corpus and approaches to entity relation extraction. We hope that BeansTALK and similar conversational question-answering tools will be helpful for the average user in an era of ongoing rapid change in climate. We believe the expansion of knowledge in plant sciences can help combat these issues (Eckardt et al., 2022).

References

2024. Pick your ai poison. *Nature Machine Intelligence*, 6(10):1119–1119.

Sören Auer, Dante A. C. Barone, Cassiano Bartz, Eduardo G. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1).

Susanne C. Brink. 2024. Rise of the machines: artificial intelligence in plant science and publishing. *Trends in Plant Science*, 29(2):101–103.

Estelle Chaix, Bertrand Dubreucq, Abdelhak Fatihi, Dialekti Valsamou, Robert Bossy, Mouhamadou Ba, Louise Deléger, Pierre Zweigenbaum, Philippe Bessières, Loic Lepiniec, and Claire Nédellec. 2016. Overview of the regulatory network of plant seed development (SeeDev) task at the BioNLP shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Nancy A Eckardt, Elizabeth A Ainsworth, Rajeev N Bahuguna, Martin R Broadley, Wolfgang Busch, Nicholas C Carpita, Gabriel Castrillo, Joanne Chory, Lee R DeHaan, Carlos M Duarte, Amelia Henry, S V Krishna Jagadish, Jane A Langdale, Andrew D B Leakey, James C Liao, Kuan-Jen Lu, Maureen C McCann, John K McKay, Damaris A Odeny, Eder Jorge de Oliveira, J Damien Platten, Ismail Rabbi, Ellen Youngsoo Rim, Pamela C Ronald, David E Salt, Alexandra M Shigenaga, Ertao Wang, Marnin Wolfe, and Xiaowei Zhang. 2022. Climate change challenges, plant science solutions. *The Plant Cell*, 35(1):24–66.

Xuejie Hao, Zheng Ji, Xiuhong Li, Lizeyan Yin, Lu Liu, Meiying Sun, Qiang Liu, and Rongjin Yang. 2021. Construction and application of a knowledge graph. *Remote Sensing*, 13(13):2511.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19. ACM.

Gaganpreet Jhajj, Xiaokun Zhang, Jerry Ryan Gustafson, Fuhua Lin, and Michael Pin-Chuan Lin. 2024. *Educational Knowledge Graph Creation and Augmentation via LLMs*, page 292–304. Springer Nature Switzerland.

M.R. Kabir and F. Lin. 2023. An llm-powered adaptive practicing system. In *AIED 2023 workshop on Empowering Education with LLMs-the Next-Gen Interface and Content Generation*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Alka Malik, Nidhi Malik, and Anshul Bhatia. 2024. *Improving Knowledge Representation Using Knowledge Graphs: Tools and Techniques*, page 381–396. Springer Nature Switzerland.

Mariana Neves and Ulf Leser. 2015. Question answering for biology. *Methods*, 74:36–46.

Bolanle Ojokoh and Emmanuel Adebisi. 2019. A review of question answering systems. *Journal of Web Engineering*, 17(8):717–758.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint*.

Andrea Papaluca, Daniel Krefl, Sergio Rodríguez Méndez, Artem Lensky, and Hanna Suominen. 2024. Zero- and few-shots knowledge graph triplet extraction with large language models. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 12–23, Bangkok, Thailand. Association for Computational Linguistics.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8.

Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Dragan Gašević, Geber Ramalho, and Rafael Ferreira Mello. 2024. Assessing the quality of automatic-generated short answers using gpt-4. *Computers and Education: Artificial Intelligence*, 7:100248.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. 2023. Enhancing knowledge graph construction using large language models. *arXiv preprint*.

Junwei Yang, Hanwen Xu, Srbuhi Mirzoyan, Tong Chen, Zixuan Liu, Zequn Liu, Wei Ju, Luchen Liu, Zhiping Xiao, Ming Zhang, and Sheng Wang. 2024. Poisoning medical knowledge using large language models. *Nature Machine Intelligence*, 6(10):1156–1168.

Yujia Zhang, Tyler Sadler, Mohammad Reza Taesiri, Wenjie Xu, and Marek Reformat. 2024. Fine-tuning language models for triple extraction with data augmentation. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 116–124, Bangkok, Thailand. Association for Computational Linguistics.

A Paper Dataset

Table 2 below shows the papers that constitute the corpus for this research.

Article Name	DOI	Edition
Deterministic responses of biodiversity to climate change through exotic species invasions	10.1038/s41477-024-01797-7	October 2024
Boreal tree species diversity increases with global warming but is reversed by extremes	10.1038/s41477-024-01794-w	October 2024
The Arabidopsis U1 snRNP regulates mRNA 3′-end processing	10.1038/s41477-024-01796-8	October 2024
Retrotransposon addiction promotes centromere function via epigenetically activated small RNAs	10.1038/s41477-024-01773-1	September 2024
Comparisons of two receptor-MAPK pathways in a single cell-type reveal mechanisms of signalling specificity	10.1038/s41477-024-01768-y	September 2024
Roles of microbiota in autoimmunity in Arabidopsis leaves	10.1038/s41477-024-01779-9	September 2024
Recruitment of Cdc48 to chloroplasts by a UBX-domain protein in chloroplast-associated protein degradation	10.1038/s41477-024-01769-x	September 2024
Enhancers associated with unstable RNAs are rare in plants	10.1038/s41477-024-01741-9	August 2024
Removal of the large inverted repeat from the plastid genome reveals gene dosage effects and leads to increased genome copy number	10.1038/s41477-024-01709-9	June 2024
Plant height as an indicator for alpine carbon sequestration and ecosystem response to warming	10.1038/s41477-024-01705-z	June 2024
The wheat powdery mildew resistance gene Pm4 also confers resistance to wheat blast	10.1038/s41477-024-01718-8	June 2024
Parental conflict driven regulation of endosperm cellularization by a family of Auxin Response Factors	10.1038/s41477-024-01706-y	June 2024
Genome resources for three modern cotton lines guide future breeding efforts	10.1038/s41477-024-01713-z	June 2024
A framework for tracing timber following the Ukraine invasion	10.1038/s41477-024-01648-5	March 2024
A scoping review on tools and methods for trait prioritization in crop breeding programmes	10.1038/s41477-024-01639-6	March 2024
Meiotic recombination dynamics in plants with repeat-based holocentromeres shed light on the primary drivers of crossover patterning	10.1038/s41477-024-01625-y	March 2024
A self-regulatory cell-wall-sensing module at cell edges controls plant growth	10.1038/s41477-024-01629-8	March 2024
O-glycosylation of the transcription factor SPATULA promotes style development in Arabidopsis	10.1038/s41477-023-01617-4	February 2024

Article Name	DOI	Edition
D6PK plasma membrane polarity requires a repeated CXX(X)P motif and PDK1-dependent phosphorylation	10.1038/s41477-023-01615-6	February 2024
A conserved graft formation process in Norway spruce and Arabidopsis identifies the PAT gene family as central regulators of wound healing	10.1038/s41477-023-01568-w	January 2024
Regulation of micro- and small-exon retention and other splicing processes by GRP20 for flower development	10.1038/s41477-023-01605-8	January 2024
A suberized exodermis is required for tomato drought tolerance	10.1038/s41477-023-01567-x	January 2024
Light-induced LLPS of the CRY2/SPA1/FIO1 complex regulating mRNA methylation and chlorophyll homeostasis in Arabidopsis	10.1038/s41477-023-01580-0	December 2023
Evolution of a plant growth-regulatory protein interaction specificity	10.1038/s41477-023-01556-0	December 2023
Targeted knockout of a conserved plant mitochondrial gene by genome editing	10.1038/s41477-023-01538-2	November 2023
Mutation in Polycomb repressive complex 2 gene OsFIE2 promotes asexual embryo formation in rice	10.1038/s41477-023-01536-4	November 2023
Evolution of phenotypic disparity in the plant kingdom	10.1038/s41477-023-01513-x	October 2023
Population genomics identifies genetic signatures of carrot domestication and improvement and uncovers the origin of high-carotenoid orange carrots	10.1038/s41477-023-01526-6	October 2023
Structural basis for abscisic acid efflux mediated by ABCG25 in Arabidopsis thaliana	10.1038/s41477-023-01510-0	October 2023
Environmental gradients reveal stress hubs predating plant terrestrialization	10.1038/s41477-023-01491-0	September 2023
A critical role of a eubiotic microbiota in gating proper immunocompetence in Arabidopsis	10.1038/s41477-023-01501-1	September 2023
Antigravitropic PIN polarization maintains non-vertical growth in lateral roots	10.1038/s41477-023-01478-x	September 2023
Ectopic callose deposition into woody biomass modulates the nano-architecture of macrofibrils	10.1038/s41477-023-01459-0	September 2023
Ocean current patterns drive the worldwide colonization of eelgrass (Zostera marina)	10.1038/s41477-023-01464-3	August 2023
Theoretical assessment of persistence and adaptation in weeds with complex life cycles	10.1038/s41477-023-01482-1	August 2023
Reorganization of seagrass communities in a changing climate	10.1038/s41477-023-01445-6	July 2023
Comparative phylotranscriptomics reveals ancestral and derived root nodule symbiosis programmes	10.1038/s41477-023-01441-w	July 2023
The plant nuclear lamina disassembles to regulate genome folding in stress conditions	10.1038/s41477-023-01457-2	July 2023

Article Name	DOI	Edition
Whole-mount smFISH allows combining RNA and protein quantification at cellular and subcellular resolution	10.1038/s41477-023-01442-9	July 2023
Next-generation ABACUS biosensors reveal cellular ABA dynamics driving root growth at low aerial humidity	10.1038/s41477-023-01447-4	July 2023
Structure and sucrose binding mechanism of the plant SUC1 sucrose transporter	10.1038/s41477-023-01421-0	June 2023
A Wox3-patterning module organizes planar growth in grass leaves and ligules	10.1038/s41477-023-01405-0	May 2023
Gibberellins promote polar auxin transport to regulate stem cell fate decisions in cambium	10.1038/s41477-023-01360-w	April 2023
Low-temperature and circadian signals are integrated by the sigma factor SIG5	10.1038/s41477-023-01377-1	April 2023
Tight genetic linkage of genes causing hybrid necrosis and pollinator isolation between young species	10.1038/s41477-023-01354-8	March 2023
A gene silencing screen uncovers diverse tools for targeted gene repression in Arabidopsis	10.1038/s41477-023-01362-8	March 2023
Newly identified sex chromosomes in the Sphagnum (peat moss) genome alter carbon sequestration and ecosystem dynamics	10.1038/s41477-022-01333-5	February 2023
Leaf transformation for efficient random integration and targeted genome modification in maize and sorghum	10.1038/s41477-022-01338-0	February 2023
Widely conserved AHL transcription factors are essential for NCR gene expression and nodule development in Medicago	10.1038/s41477-022-01326-4	February 2023
Economic and biophysical limits to seaweed farming for climate change mitigation	10.1038/s41477-022-01305-9	January 2023
Control of plastid inheritance by environmental and genetic factors	10.1038/s41477-022-01323-7	January 2023
Cell-type-specific PtrWOX4a and PtrVCS2 form a regulatory nexus with a histone modification system for stem cambium development in Populus trichocarpa	10.1038/s41477-022-01315-7	January 2023
Direct attenuation of Arabidopsis ERECTA signalling by a pair of U-box E3 ligases	10.1038/s41477-022-01303-x	January 2023
The eINTACT system dissects bacterial exploitation of plant osmosignalling to enhance virulence	10.1038/s41477-022-01302-y	January 2023
Cryo-EM structure of the respiratory I + III2 supercomplex from Arabidopsis thaliana at 2 Å resolution	10.1038/s41477-022-01308-6	January 2023
Plant-specific features of respiratory supercomplex I + III2 from Vigna radiata	10.1038/s41477-022-01306-8	January 2023
Dynamic chromatin accessibility deploys heterotypic cis/trans-acting factors driving stomatal cell-fate commitment	10.1038/s41477-022-01304-w	December 2022
Towards a unified theory of plant photosynthesis and hydraulics	10.1038/s41477-022-01244-5	November 2022

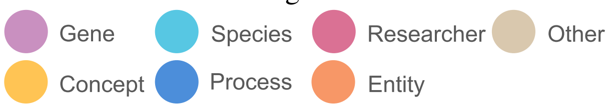
Article Name	DOI	Edition
Algal photosystem I dimer and high-resolution model of PSI-plastocyanin complex	10.1038/s41477-022-01253-4	October 2022
Dynamic genome evolution in a model fern	10.1038/s41477-022-01226-7	September 2022
Tethering of cellulose synthase to microtubules dampens mechano-induced cytoskeletal organization in Arabidopsis pavement cells	10.1038/s41477-022-01218-7	September 2022
PIF4 enhances DNA binding of CDF2 to co-regulate target gene expression and promote Arabidopsis hypocotyl cell elongation	10.1038/s41477-022-01213-y	September 2022
The slow-evolving Acorus tatarinowii genome sheds light on ancestral monocot evolution	10.1038/s41477-022-01187-x	July 2022
State of ex situ conservation of landrace groups of 25 major crops	10.1038/s41477-022-01144-8	May 2022
The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence	10.1038/s41477-022-01146-6	May 2022
Sucrose synthases are not involved in starch synthesis in Arabidopsis leaves	10.1038/s41477-022-01140-y	May 2022
Modelling the pyrenoid-based CO2-concentrating mechanism provides insights into its operating principles and a roadmap for its engineering into crops	10.1038/s41477-022-01153-7	May 2022
Genomes of leafy and leafless Platanthera orchids illuminate the evolution of mycoheterotrophy	10.1038/s41477-022-01127-9	April 2022
The Cycas genome and the early evolution of seed plants	10.1038/s41477-022-01129-7	April 2022
R-loops at microRNA encoding loci promote co-transcriptional processing of pri-miRNAs in plants	10.1038/s41477-022-01125-x	April 2022
Targeted introduction of heritable point mutations into the plant mitochondrial genome	10.1038/s41477-022-01108-y	March 2022
Carbon flux through photosynthesis and central carbon metabolism show distinct patterns between algae, C3 and C4 plants	10.1038/s41477-021-01042-5	January 2022
Representation and participation across 20 years of plant genome sequencing	10.1038/s41477-021-01031-8	December 2021
Loss-of-function alleles of ZmPLD3 cause haploid induction in maize	10.1038/s41477-021-01037-2	December 2021
An siRNA-guided ARGONAUTE protein directs RNA polymerase V to initiate DNA methylation	10.1038/s41477-021-01008-7	November 2021
The root meristem is shaped by brassinosteroid control of cell geometry	10.1038/s41477-021-01014-9	November 2021
Exogenous miRNAs induce post-transcriptional gene silencing in plants	10.1038/s41477-021-01005-w	October 2021
Insights into angiosperm evolution, floral development and chemical biosynthesis from the Aristolochia fimbriata genome	10.1038/s41477-021-00990-2	September 2021
The Taxus genome provides insights into paclitaxel biosynthesis	10.1038/s41477-021-00963-5	August 2021

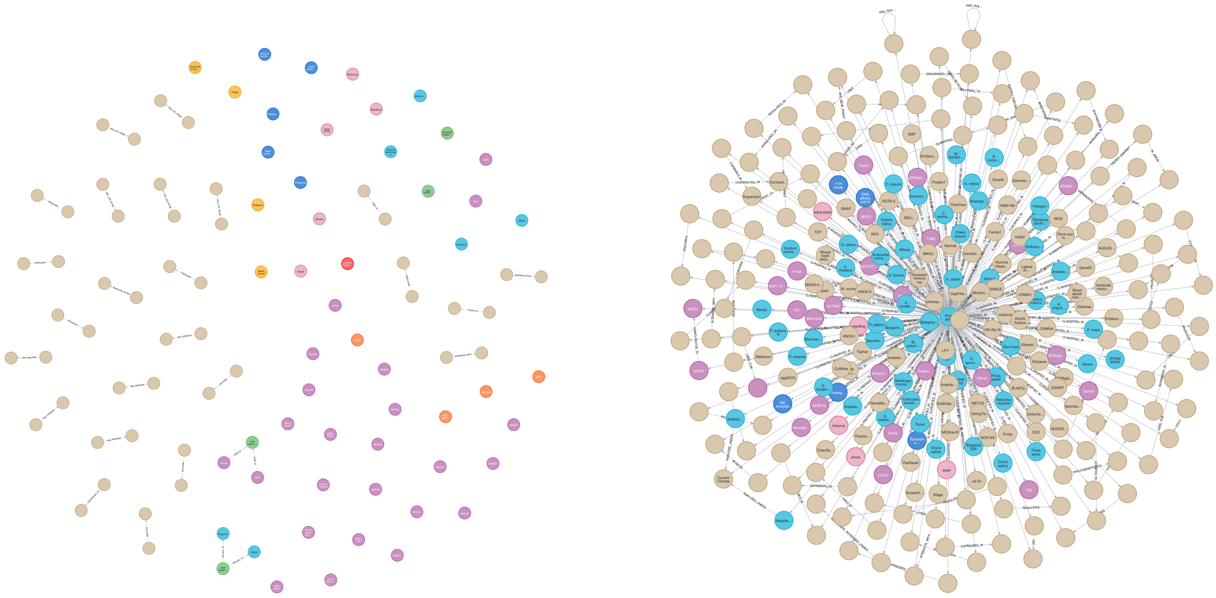
Article Name	DOI	Edition
A microbiota–root–shoot circuit favours Arabidopsis growth over defence under suboptimal light	10.1038/s41477-021-00956-4	August 2021
Coordination of microbe–host homeostasis by crosstalk with plant innate immunity	10.1038/s41477-021-00920-2	June 2021
The reference genome of Miscanthus floridulus illuminates the evolution of Saccharinae	10.1038/s41477-021-00908-y	May 2021
Molecular landscape of etioplast inner membranes in higher plants	10.1038/s41477-021-00896-z	April 2021
A scoping review of adoption of climate-resilient crops by small-scale producers in low- and middle-income countries	10.1038/s41477-020-00783-z	October 2020
A scoping review of feed interventions and livelihoods of small-scale livestock keepers	10.1038/s41477-020-00786-w	October 2020
A high-contiguity Brassica nigra genome localizes active centromeres and defines the ancestral Brassica genome	10.1038/s41477-020-0735-y	August 2020
Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts	10.1038/s41477-020-0618-2	March 2020
Genomes of early-diverging streptophyte algae shed light on plant terrestrialization	10.1038/s41477-019-0560-3	February 2020
The hornwort genome and early land plant evolution	10.1038/s41477-019-0588-4	February 2020
Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus	10.1038/s41477-019-0577-7	January 2020
Musa balbisiana genome reveals subgenome evolution and functional divergence	10.1038/s41477-019-0452-6	August 2019
Genome structure and evolution of Antirrhinum majus L	10.1038/s41477-018-0349-9	February 2019
Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution	10.1038/s41477-018-0337-0	January 2019
Fern genomes elucidate land plant evolution and cyanobacterial symbioses	10.1038/s41477-018-0188-8	July 2018
A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits	10.1038/s41477-018-0166-1	July 2018
A genome for gnetophytes and early evolution of seed plants	10.1038/s41477-017-0097-2	February 2018
The Aegilops tauschii genome reveals multiple impacts of transposons	10.1038/s41477-017-0067-8	December 2017
The rubber tree genome reveals new insights into rubber production and species adaptation	10.1038/nplants.2016.73	June 2016
Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida	10.1038/nplants.2016.74	June 2016

Table 2: Nature Plants open access papers that constitute the corpus for this preliminary work.

B Knowledge Graphs Generated

Many different tool options exist for creating KGs (Malik et al., 2024). In this work, we leveraged Neo4j for KG modeling as it is a popular choice as an open-source graph database (Hao et al., 2021). The legend below shows the labels for the entities in the KGs generated for the six most common entity types.





(a) GPT-4

(b) GPT-4 Turbo

(c) GPT-4o

(d) GPT-4o mini

Figure 3: KGs generated by four different models.

C Example Cypher Code

A truncated example summary of the LLM suggested nodes for one of the 99 papers. The below output was created with GPT-4o on the paper titled *The Arabidopsis U1 snRNP regulates mRNA 3′-end processing*.

```
// Create Nodes
// Create species node
CREATE (a:Species {name: "Arabidopsis␣thaliana"});

// Create gene/protein nodes
CREATE (e1:Protein {name: "EIF4A-III"}),
       (e2:Protein {name: "RS2Z33"}),
       (e3:Protein {name: "GPX6"}),
       (e4:Protein {name: "At1g13690"}),
       (e5:Protein {name: "IMPA-2"}),
       (e6:Protein {name: "CHC1"}),
       (e7:Protein {name: "CRTISO"}),
       (e8:Protein {name: "ATRZ-1A"}),
       (e9:Protein {name: "AT5G55670"}),
       (e10:Protein {name: "MDAR1"}),
       (e11:Protein {name: "At2g36400"}),
       (e12:Protein {name: "SC35"}),
       (e13:Protein {name: "At2g06200"}),
       (e14:Protein {name: "AT2G25970"}),
       (e15:Protein {name: "SGS3"}),
       (e16:Protein {name: "SPT16"}),
       (e17:Protein {name: "AT5G51410"}),
       (e18:Protein {name: "AT1G60900"}),
       (e19:Protein {name: "HTB9"}),
       (e20:Protein {name: "MTA"}),
       (e21:Protein {name: "AT1G33680"}),
       (e22:Protein {name: "HSC70-1"}),
       (e23:Protein {name: "MTHFR1"}),
       (e24:Protein {name: "emb1220"}),
       (e25:Protein {name: "NRPB4"}),
       (e26:Protein {name: "AT2G34040"}),
       (e27:Protein {name: "AT5G53440"}),
       (e28:Protein {name: "Y14"}),
       (e29:Protein {name: "AT2G18740"}),

// Create molecule nodes
CREATE (m1:Molecule {name: "GFP"}),
       (m2:Molecule {name: "Myc-CBP20"}),
       (m3:Molecule {name: "YFP-RBP47B"}),
       (m4:Molecule {name: "RFP"});

// Create interaction relationships
MATCH (m1:Molecule {name: "GFP"}), (p:Protein {name: "U1␣snRNP␣components"}), (m2:Molecule {name:
     "Myc-CBP20"})
CREATE (p)-[:INTERACTS_WITH]->(m2);

MATCH (m4:Molecule {name: "RFP"}), (m3:Molecule {name: "YFP-RBP47B"})
CREATE (m4)-[:INTERACTS_WITH]->(m3);
```

Future efforts in this area must explore the scientific validation of extracted entities and relationships. As the volume of AI-generated data rapidly increases, it will become more important and challenging to implement validation processes to prevent misinformation from impacting critical knowledge bases (nat, 2024). Work has shown that even minute amounts of AI-generated errors can corrupt biomedical KGs (Yang et al., 2024).