HYPOTHESES FOR A NEW GENERATION: LEVERAGING NATURAL LANGUAGE PROCESSING TO BRIDGE GAPS AND GENERATE NOVEL HYPOTHESES FOR DESICCATION TOLERANCE RESEARCH

By

Serena Ghantous Lotreck

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Biology—Doctor of Philosophy
Molecular Plant Sciences—Dual Major

2024

ABSTRACT

Scientific hypotheses, which are explanations of natural phenomena that can be tested and falsified, are at the core of empirical biology research. Hypotheses about genes involved in biological processes or interactions between species in an ecological setting are used to design research studies and make discoveries about the natural world. However, the act of generating a novel hypothesis requires a high level of manual labor, including sifting through and reading numerous previously published research articles. Due to the explosion of scientific literature in the last century, there are too many materials in any given field for scientists to read and process while generating new hypotheses, leading to a sensation of information overload. Information overload is the state when information inputs to a system overwhelm its information processing capacities, and is not a new phenomenon; since the advent of the written word, academics have bemoaned the deluge of written resources. One possible method for ameliorating the sensation of information overload is to implement methods for automated hypothesis generation, whereby literature is automatically processed to propose new connections between biological entities. In particular, this dissertation focuses on the use of knowledge graphs, which are networks in which nodes are entities of interest, like genes or proteins, and edges are the biological relationships between them. While methods for automated hypothesis generation from the literature using knowledge graphs have been used in the biomedical literature to generate hypotheses for phenomena like adverse drug reactions or drug-disease interactions, limited work has been done to translate these methods into the plant science domain.

This dissertation focuses on the use of natural language processing techniques to perform automated hypothesis generation in and explore the research landscape of the field of desiccation tolerance biology. Desiccation tolerance is the ability of an organism to revive from the loss of nearly all internal water, and exists across the kingdom of life. Nearly all land plants exhibit desiccation tolerance in seeds; however, whole-plant vegetative desiccation tolerance is much rarer, and whole-organism desiccation tolerance in other kingdoms of life is also rare. As a result, the field of desiccation tolerance research is much smaller than related fields such as drought

tolerance, and possesses many fewer curated resources both experimentally, like transformation systems for desiccation tolerant organisms, as well as informationally, as manually curated databases focus on model and crop species which do not exhibit whole-plant desiccation tolerance. Many current knowledge graphs in the plant sciences are built from manually curated databases such as Planteome and UniProt, and are therefore lacking rich information on desiccation tolerance from which to generate hypotheses. Automatic information extraction from the scientific literature to identify new entities and relationships in an understudied group of organisms in a high-throughput manner is therefore promising as an approach to ameliorate the data gaps in databases that affect knowledge graph-based hypothesis generation. The first chapter of this dissertation reviews the history of information overload and hypothesis generation, and briefly introduces desiccation tolerance as a research system. Chapter two presents a dataset for the molecular plant sciences labeled with biological entities and relationships that can be used to train information extraction models, and evaluate several existing methods on this dataset. In chapter two, I find that models from other scientific disciplines are insufficient for high-quality information extraction in plant science, and that training a new model yields improved performance. In chapter three of this thesis, I use bibliometric methods and topic modeling to explore the research landscape of desiccation tolerance, and find that the various study systems (animal, plant, fungi and microbe) are very siloed, or isolated, from one another, even though mechanisms for desiccation tolerance are shared across the kingdoms of life. Additionally, I design a rule-based algorithm to use bibliometric data to recommend new attendees to a specialized desiccation tolerance conference. Finally, in the fourth chapter, I explore the possibilities for constructing a knowledge graph of desiccation and drought tolerance research, and of using the resulting graph to predict novel hypotheses about the mechanisms of desiccation tolerance. My work shows that, using the chosen data sources and methods, information extraction and hypothesis generation from knowledge graphs are inadequate to generate high-quality hypotheses. In the final chapter, I reflect on the limitations and potential future directions of automated hypothesis generation for biology. This research will hopefully provide insight on information management and hypothesis generation in the plant sciences.

ACKNOWLEDGEMENTS

The last five years have been nothing at all like what I expected them to be, in so many ways. While I've wanted to earn a PhD since my senior year of high school, nothing could have prepared me for the highs and lows of the experience itself, and I have so many people to thank for helping me see this crazy venture through.

As a freshman in undergrad, I toured the Jander Lab through the Cornell Undergraduate Research Board's peer mentorship program. Cairo, my peer mentor who ran the tour, asked one of the lab's postdocs to accompany us, and I will never forget what this postdoc replied when one of the students asked her why she chose to work in plant biology: "I like working on plants because they're interesting, but mostly, I stay because the people are so nice."

A year later, I joined the Jander Lab as an undergraduate research intern, and that postdoc's words have rang true for me every day since. The plant scientists that I have been fortunate enough to meet and work with over the last 9 years of my education have shaped me and helped form my career trajectory as a researcher, and the kindness and mentorship that they've shown me has been formative in how I operate as a professional and a researcher. In addition to the plant scientists, there are so many other people that were instrumental in my development as a scientist and a person over the last five years, and they all certainly deserve more thanks than I am able to put into words here, but I will do my utmost!

From my undergraduate, I'd like to thank Dr. Tom Silva, the lecturer in my freshman physiology course that first made me excited about plants, and Cairo, my peer mentor, for opening the doors to undergraduate research in plant science. To Dr. Georg Jander for accepting me into his lab, and Kevin Ahern for his considerate and personable mentorship style as he taught me the basics of being a scientist, and for his continued friendship as we've both pursued our PhDs. I also owe a large thanks to Dr. Suman Seth, for always pushing me to take a productively critical view of science through the lens of history, and to Andrew, Mark and Karel and all the other people I worked with at Cornell Outdoor Education for giving me opportunities to be a leader and push myself to become a more capable person.

At MSU, I'd like to extend a huge thank you to my advisors, Drs. Bob VanBuren and Mohammad Ghassemi. A special thanks to Bob for his role as my REU program mentor during my junior year of undergrad, as it piqued my interest in computational science, setting the stage for the career transition to data science that I have made as a graduate student. Thank you to the both of you for your continued mentorship and guidance as I've worked towards finishing this dissertation, and thank you for supporting my research interests, even though they didn't fit neatly into either of your research programs. Thanks as well to my committee, Drs. Tammy Long, Emily Josephs, and Dan Chitwood, for your feedback and guidance.

To everyone in the VanBuren lab, thank you all for embracing me as your labmate. Your encouragement, friendship, support, and of course, the tea, has brightened so many of my days while finishing this degree. You have made my daily work environment such a happy and healthy one, and I am privileged to call you all my coworkers and friends.

To Sara Lira at Corteva, thank you for having faith in me and taking me on as your intern. Working with you has been a privilege and a joy, and your mentorship over the last several years has been instrumental in my success in my degree and beyond.

To my Plant Biology program cohort, past labmates and other MSU friends, thank you all for the brunches, escape rooms, hikes, and porch beers over the last five years. Even during Deep Covid, your willingness to participate in virtual game nights or outdoor meetups was such a blessing. As we've all started to graduate and move away, I am profoundly grateful to all of you for making my time here so wonderful.

To the office staff in Plant Biology, especially Sara Kraeuter; without your support I would have missed so many logistical milestones in this program. Thank you for always being so responsive and helpful, even for the silliest of questions! To the developers I've interacted with in the last five years, especially Dave Wadden, Harry Caufield and Max Berrendorf, thank you for your technical support and willingness to answer questions and help me implement your codebases for my research; I'm a better programmer and scientist for our interactions.

To my roommates, past and present. To Cass, por nuestras aventuras mientras estabas aquí y las

llamadas semanales desde que te fuiste, en que me impartas siempre los mejores chismes y consejos buenos. To Nick and Jacob (and honorary 4th roommate Julia!) for putting up with me during this last stretch of thesis work, and for always being there with a, "CoDe GeAHsS??" when I needed a distraction. A special thanks to Nick, my longest-standing roommate, for your unwavering support during some of the most difficult times in my PhD. Our tea-and-Ted Lasso nights were the light of my third year, and I am so eternally grateful for your friendship.

To everyone at the Greater Lansing Academy of Dance for welcoming me with open arms and making me a part of your community. As I like to call it, my "enforced fun time" of classes, rehearsals, and subsequent parking lot hangouts was many times the only thing keeping me from losing my mind over my dissertation work in the last few years. Having such a wonderful, fun community outside of school was essential to my well-being, and I so very much wish I could bring you all with me wherever I end up next! A special thanks to my instructor Jim, for pushing me to believe that I am capable of improvement even as an adult, and giving me an outlet to work towards something productive outside of research. Your thoughtful observation and explanations have made me so much stronger of a dancer, and I have so enjoyed working with you.

A Migue, Alberto, Angel, y David, por acogerme como parte del grupo desde el principio, desde las llamadas de los domingo durante la pandemia hasta las quedadas de este año; sois unos reyes del Martes Santo! A Barto Miranda, por tus años de amistad y todas nuestras charlas y paseos; siempre me trae mucha alegría pasar tiempo contigo. Y finalmente a Martín, por ser tan buen amigo mío por tantos años despues de conocerme por casualidad en el roco. Tenerte siempre allí para hablar de todo lo que pasa en nuestras vidas, tanto personal como de investigación, y seguir teniendo aventuras contigo me alegra más de lo que puedo explicar.

I would like to extend a heartfelt thanks to all of the artists and production professionals that are responsible for making the music, movies, and shows that helped get me through the toughest times of my doctoral degree. I would not have been able to accomplish meaningful scientific work without the background of movie soundtracks and pop songs that accompanied my programming, reading, writing, and analysis, to say nothing of the movies, books, shows, and podcasts that

enriched my world in my free time. Thanks as well to the therapists and medical professionals that have supported my journey; graduate school is hard on the mind and the body, and without the support of the medical and mental health professionals I've been fortunate to work with over the years, I wouldn't be here today.

To my best friend Galen, I can't express in words how much your support and friendship has meant to me. You push me to be a better person every day, and our phone calls and road trips and general shenanigans have made my world so much brighter since we met that fateful day in Okenshields so long ago.

Finally, to my family. To my Grandma, for being my best friend and confidante on all of our phone calls, and for sending me newspaper clippings and notes to brighten my spirits. To my siblings Robert and Jake and my parents: while we all hope there are no more world-halting pandemics in our lifetimes, I am so profoundly grateful that I got locked in with you all in 2020. Living with you while all being adults was such a rare and wonderful opportunity, and the board games and shared meals, spontaneous kitchen dance parties and outdoor activities are all such cherished memories to me. Your support and understanding of the trials and tribulations of this degree and everything that happens in my life, both in person and over distance, have been so important to me over the last, well, my whole life. I love you all so much.

PREFACE

"What if I told you you'd never have to read a scientific paper again?" As an undergraduate student, the proposal for a dissertation project on automated hypothesis generation sounded like a proposal for the promised land. I had already been burned scientifically by not reading enough papers; while writing up my undergraduate honors thesis research, I found a paper that, had I read it earlier, would have drastically changed my experimental design. Ironically, I have of course read more papers to complete the project described here by several orders of magnitude than for any other project I have worked on, and have continued to experience the same phenomenon of being unfortunately surprised by relevant papers appearing at the wrong moments. However, the research and writing of this thesis has assured me that it is not as a result of some deficiency as a scientist, but is rather an eternal struggle that has existed since the advent of the written word.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

TLDR	too long; didn't read
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
PO	Plant Ontology
PECO	Plant Experimental Conditions Ontology
KG	Knowledge Graph
NLP	Natural Language Processing
NER	Named Entity Recognition
RE	Relation Extraction
PICKLE	Plant Science Knowledge Graph (Corpus)
LP	Link prediction
TLP	Temporal link prediction
AUROC	Area under the receiver-operator curve
ROC	Receiver-operator characteristic
AUPRC	Area under the precision-recall curve

CHAPTER 1

INTRODUCTION

Information overload

Information overload is the state when information inputs to a system overwhelm the system's information processing capacities [Bawden and Robinson, 2020]. One can experience an intuitive example of information overload simply by using an academic search engine to look up a concept in one's research field; the hundreds of thousands of results serve as a testament to the sheer amount of information available even in a relatively narrow scope. Information overload is often perceived to be associated only with the modern digital age, as a result of the advent of information technologies like the Internet, but humankind has been complaining of information overload, and developing strategies to deal with it, for nearly as long as we have had written text. In the first century A.D., the Roman philosopher Seneca griped that "the abundance of books is distraction" [Bawden and Robinson, 2020]. Vincent of Beauvais, a Christian academic who wrote compendiums of available knowledge in the mid-13th century (a strategy for the management of information overload even in an era before the advent of the printing press), bemoaned "the multitude of books, the shortness of time, and slipperiness of memory" [Bawden and Robinson, 2020] – a complaint that, when replacing "books" with "journal articles", I have found wholly relatable in the writing of this dissertation! Information overload exists across all spheres of life where written information dominates; however, concern over the effect of information overload on the future progress of science is acute [Raymond, 2019]. Scientists rely upon previous information to generate hypotheses and design experiments to make scientific discoveries, but we are unable to keep up with the flow of information even within very specific domains, often relying on information management approaches that involve manually parsing, reading, and digesting the resulting papers to formulate new hypotheses [Landhuis, 2016].

Given that a perception of overload has existed since humankind started writing things down, it is unlikely that we will ever manage to design a "silver bullet" tool or set of tools that so effectively manages our information workflows that the perception of being overloaded recedes substantially.

However, while they have not necessarily lessened our perception of overload, previous strategies for information management, like Vincent de Beauvis' encyclopedic compendium *Speculum Maius* or the Dewey Decimal system, have still proved fruitful. Without attempting to manage overflow, our ability to navigate available information in any period of time would be deeply hampered, and it is therefore imperative to continue to develop new strategies to maintain pace with humanity's ever-expanding body of knowledge. Indeed, some authors argue that our contemporary perception of information overload is more related to the lack of technological solutions for managing digital information than it is to the existence of large quantities of information itself [Klerings et al., 2015].

Existing tools to manage information overload in the plant sciences

A range of digital tools exist to help manage information overload in the sciences, ranging from familiar approaches like search engines to other newer approaches, like knowledge graphs. One excellent example of a domain-agnostic search engine-based tool for information management is Semantic Scholar [Raymond, 2019], which has incorporated various machine learning approaches to search retrieval and information management. For example, in 2020 Semantic Scholar incorporated TLDRs ("too long; didn't read"'s), which are one- to two-sentence summaries of scientific abstracts, using a machine learning model for extreme summarization [Cachola et al., 2020]. The goal of TLDRs is to assist scientists in identifying relevant papers from a search more rapidly than possible by reading entire abstracts. In addition to domain-agnostic tools, plant science researchers have access to a relatively large number of high-quality, manually-curated databases. Some are specific to plant science, like Planteome [Cooper et al., 2024], while others are generalizable to all areas of biology, like the Gene Ontology (GO) [The Gene Ontology Consortium, 2019] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa, 2002]. Planteome in particular provides a valuable service to plant scientists, specializing in developing ontologies specific to plant science like the Plant Ontology (PO) and the Plant Experimental Conditions Ontology (PECO) and mapping information from the plant science literature onto these ontologies to make databases, as well as linking other ontology projects. While databases and ontologies provide extremely high-quality information, their scope is limited by the labor that needs to be invested in manual

curation.

A more recent approach to information management in the sciences is knowledge graphs. A knowledge graph (KG) is a network that contains knowledge of the real world, where nodes in the graph are entities of interest, and edges are relations between the entities [Peng et al., 2023]. In the biological sphere, the capacity of KG to contain heterogeneous information, or information from various sources and of various types, has made them attractive candidates for combining various ontologies and databases for tasks such as predicting new links between biological entities [Unni et al., 2022]. Knowledge graphs can include information from both structured and unstructured sources, meaning they can integrate information from across structured databases with information extracted directly from the scientific literature, patents, or other forms of unstructured natural language text. KG themselves cannot solve the issue of information overload; a KG that represents some large amount of information is also intractably large. However, different from search engine-based approaches, KG can be used in downstream methods that aim to manage information in a very specific way: by automating or semi-automating the creation of scientific hypotheses.

What is a hypothesis?

Before we can discuss the automation of hypothesis generation, we need to establish a definition for a hypothesis. In this work, we will define a hypothesis following [Alger, 2019]: as a proposed explanation of a natural phenomenon that can be tested and potentially falsified. A hypothesis is a "putative explanation for actual observations" [Alger, 2019] from which predictions about the behavior of a system can be derived. For example, we might observe that our houseplants have been turning yellow, and hypothesize that the reason is a lack of nutrients in the soil. One prediction resulting from this hypothesis is that if we added nutrients to the soil by adding fertilizer, our plants would become green again. We can evaluate this prediction by performing the experiment of adding fertilizer, and if our plants do not turn green, we can reject the hypothesis of nutrient deficiency as a cause for plant yellowing. [Alger, 2019] notes that this framing of the scientific process as the falsification of hypotheses can be controversial among scientists, some of whom

argue for open discovery- or question-based science as opposed to hypothesis-driven science. The full nuance of this debate is beyond the scope of this dissertation; however, I would like to point out that discovery- and hypothesis-based approaches to science appear to be synergistically integrated in the pursuit of managing information overload. This dissertation is framed around the pursuit of a hypothesis generation system for the plant sciences. Such a system can function as an open discovery- or question-based system, broadly searching within the discipline of plant science for explanations of natural phenomena; those explanations are hypotheses, which can then be experimentally tested with falsification. [Alger, 2019] defines 6 characteristics of a good hypothesis: (1) Significance/Generality, (2) Riskiness, (3) Simplicity, (4) Specificity, (5) Constraint, and (6) Falsifiability in Practice. In order, these principles require a hypothesis to (1) tackle a scientifically meaningful issue, (2) make non-obvious predictions, (3) provide the simplest explanation for the observed facts (i.e. follow Occam's Razor), (4) rule out other explanations, such as hypothesizing "necessary and sufficient" conditions in biochemistry, (5) be sufficiently detailed such that changing any detail of the hypothesis means it no longer explains what it was intended to explain, and (6) it can be falsified by experiments that are practical to perform. A good automated hypothesis generation system will generate good hypotheses; we can use these 6 characteristics to define what we mean by good hypotheses.

History of automated hypothesis generation

Now that we have a definition of hypothesis from which to work, we can define the practice of automated hypothesis generation. Here, we will define automated hypothesis generation as the practice of using an algorithmic system to propose falsifiable hypotheses for a given scientific domain. This is in contrast to manual hypothesis generation, like the process of generating a hypothesis about our dying houseplant: in that case, we used our previous knowledge about plants, which we could have gained from interpersonal interactions (talking to our houseplant-mom friend) or from reading (literature or the general Web), to come up with a plausible explanation for what we observed. In a scientific domain, manual hypothesis generation often involves reading multitudes of journal articles in the target domain to acquire sufficient intuition to generate a

hypothesis [Akujuobi, 2021]. Most literature on automated hypothesis generation from scientific papers credits the development of the field to Don R. Swanson. Swanson's major contribution to the field of automated hypothesis generation from literature (what he called literature-based discovery, or LBD) was the idea of "undiscovered public knowledge", and the ABC model approach to discovering this knowledge. Undiscovered public knowledge is information that has been implicitly demonstrated in sources like scientific papers, but has parts that have never explicitly been brought together to state that knowledge explicitly. Undiscovered public knowledge can best be explained with the example that Swanson used in his seminal paper on using fish oil to treat Raynaud's disease [Swanson, 1986], where he used the ABC model to make implicit knowledge, explicit. In the ABC model, the user provides two terms, A and C, that they think may be connected, and the system, whether it be automated or manual, searches for terms B that bridge the gap between A and C [Smalheiser and Swanson, 1998]. In [Swanson, 1986], Swanson used the ABC technique to demonstrate that, while the scientific community knew that fish oil helped improve blood flow, and that Raynaud's disease was caused by poor blood flow, no one had postulated whether, or how, consuming fish oil contributed to easing the symptoms of Raynaud's disease. In the ABC model, A in this case would be "fish oil", and C would be "amelioration of Raynaud's disease", while B is the mechanisms by which fish oil could contribute to the amelioration of Raynaud's disease. Fish oil being a treatment for Raynaud's disease is a prime example of undiscovered public knowledge; all of the information necessary to make the conclusion was present in the literature, but due to disciplinary siloing of research, had never explicitly been brought together. Other examples of undiscovered public knowledge include information on genetics that lies hidden in public databases [Smalheiser and Swanson, 1998], as well as information that is implicit in the literature.

While powerful, the ABC approach still requires some initial level of hypothesis, as discussed in [Smalheiser, 2012]. In Swanson's research, he relied on a closed paradigm, having chosen the A and C terms in advance, and looked for B terms that connected them; the selection of both an A and C term requires knowledge of the research field and an initial hypothesis about what A's and C's may be connected. While open-discovery ABC models, where only an A term needs to

be specified, exist from a technical standpoint, they result in a deluge of potential connected B and C terms, which induces further information overload that needs to be managed by ranking the resulting hypothesis candidates [Wren, 2008].

One approach to open discovery-based hypothesis generation that is somewhat more constrained than an open ABC approach is link prediction on KG. In the biomedical sphere, KG have been combined with prediction techniques to predict adverse drug interactions, new targets for drug repurposing, and drug discovery [Abu-Salih et al., 2023]. There exist a fair number of plant science KG centered around model species. AgroLD [Larmande and Todorov, 2021] is a plant science knowledge graph built from other biological databases such as UniProtKB, GO and genetic database resources for several plant species. KnetMiner [Hassani-Pak et al., 2021] is a commercialized KG platform that integrates information from genome annotations for various model species, as well as single nucleotide polymorphism variation, quantitative trait loci, and protein domains [Hassani-Pak et al., 2016]. KnetMiner does include information derived from PubMed abstracts; however, it is unclear how this information was extracted for inclusion into KnetMiner [Hassani-Pak et al., 2016]. In 2023, there was a small burst of new papers published on plant science-specific KG, partially aided by the journal Frontiers' special edition, *Knowledge Graph Technologies: the Next Frontier of the Food, Agriculture, and Water Domains*. One graph from the Frontiers edition is GenoPhenoEnvo, a graph integrated data from Planteome [Cooper et al., 2024], including both ontology information as well as gene expression data for several model and crop species [Thessen et al., 2023]. C3P0 is another KG from the Fronteirs special edition designed for providing decision support to vegetable farmers, that is built on existing databases as well as informational input from domain experts [Darnala et al., 2023]. Finally from the Frontiers group is OrthoLegKB, which directly uses genomic resources to compute and include orthology and synteny, QTL's, and RNA-sequencing datasets [Imbert et al., 2023]. Other plant science KG published in 2023 are: PlantConnectome , which used a GPT-based approach to turn 100,000 plant biology abstracts into a KG with a navigable GUI component, allowing users to explore various subsets of the graph [Fo et al., 2023]; and The Comprehensive Knowledge Network (part of the

Stress Knowledge Map, [Bleker et al., 2023]), which is a manually-curated network of *Arabidopsis thaliana* genes, proteins, RNA, and metabolites derived from literature [Ramšak et al., 2018].

However, it appears that, in comparison with the biomedical domain, very little work has been done on downstream methods for using any of these plant science KG to predict new hypotheses. Most of the better-developed KG include some kind of browser that allows users to interact with information within the graph in response to some search query; however, link prediction to generate novel hypotheses does not appear to have been investigated in the plant sciences. KG clearly provide an advantage when accessing omic-scale information to identify gene targets, as demonstrated by the search queries implemented in [Thessen et al., 2023] and [Imbert et al., 2023]; however, the same searches and results could likely have been performed, albeit with greater difficulty, using biological datasets directly, and are not unique to the KG. This is in contrast to work in the biomedical sphere that has directly utilized predictions of new graph connections to answer questions that were not answerable with other kinds of data [Abu-Salih et al., 2023]. In contrast, C3PO provides a decision-support framework, which allows farmers to input their farm details and receive a tailored technical itinerary for their planting season [Darnala et al., 2023]. While this is much closer to the kind of hypothesis generation we are interested in, it is targeted at a practical use case and not at basic biological discovery.

Desiccation tolerance as a biological system

The biological system on which this dissertation focuses is whole-plant desiccation tolerance. Desiccation tolerance (DT) is defined as the ability to revive from the "air-dry state", where all available water in the organism has been lost to the surrounding air [Bewley, 1979]. As water is the primary ingredient of life, that any organism can survive through near-complete drying is astounding, and understanding the mechanisms by which this phenomenon is possible is of great scientific interest [Hibshman et al., 2020]. Many land plants have desiccation tolerant seeds (also known as orthodox seeds), but whole-organism DT is much rarer. It is thought that the earliest land plants had whole-organism DT, which was then lost as plants evolved vasculature (xylem and phloem, which allow the long-distance internal transport of water and sugars), and that

certain plants re-evolved the trait by repurposing seed DT mechanisms under certain evolutionary pressures [Marks et al., 2021]. While DT in whole organisms is a relatively rare phenomenon, it exists across all kingdoms of life [Alpert, 2005]; animals such as tardigrades are desiccation tolerant [Hibshman et al., 2020], as well as many microbes [Grzyb and Skłodowska, 2022]. As such, the biology of DT has applications across many fields, including medical cryopreservation and crop improvements [Alpert, 2005], space biology [Persson et al., 2011], and restoration ecology [León-Lobos et al., 2012].

Because DT in whole organisms is rather rare, the field of DT research is relatively small compared to related disciplines like drought tolerance. In plant science, no model or crop organisms exhibit vegetative DT, and experimentally validated information about the mechanisms of DT is scarce. As a result, using search terms like "desiccation" or "desiccation tolerance" in large KG like AgroLD or KnetMiner returns very few results, none of which I have found to go beyond established knowledge about DT mechanisms. Therefore, the overarching goal of this dissertation is to explore the potential of using the DT literature to construct and generate hypotheses about whole-plant DT.

Content roadmap

In this dissertation, I explore the application of natural language processing to KG construction from and characterization of the DT literature. In Chapter 2, I establish the creation of a molecular plant science dataset of 250 abstracts labeled with biological entities like genes, organisms and proteins, and the relations between them, and use it to demonstrate the performance of existing methods for entity and relation extraction in the plant sciences. In Chapter 3, I explore the research themes present in the DT literature, and characterize the extent of siloing between the research in plant, animal, microbial and fungal DT research, and address these citation gaps by designing an algorithm to increase research integration through recommending new attendees to a specialized DT conference. In Chapter 4, I explore the potential of literature-derived KG to predict novel hypotheses in plant vegetative DT. Finally, in Chapter 5, I reflect on the limitations of literature-derived KG and propose future directions.

CHAPTER 2

PLANT SCIENCE KNOWLEDGE GRAPH CORPUS: A GOLD STANDARD ENTITY AND RELATION CORPUS FOR THE MOLECULAR PLANT SCIENCES

The work in this chapter is presented in the final publication:

Lotreck, S., Segura Abá, K., Lehti-Shiu, M. D., Seeger, A., Brown, B. N. I., Ranaweera, T., Schumacher, A., Ghassemi, M., and Shiu, S.-H. (2023). Plant Science Knowledge Graph Corpus: a gold standard entity and relation corpus for the molecular plant sciences. *in silico Plants*, 6(1):diad021

Author contributions:

S.L. and S.H.S. developed the project idea. S.L. designed the ontologies and annotation guidelines and wrote code to collect abstracts, unify annotations, apply and evaluate models, and create figures and manually reviewed and unified all abstracts and wrote the initial draft and figure legends. K.S.A. contributed to analyses of unexpected model performance. K.S.A., M.L.S., A.S., B.B., T.R. and A.S. annotated abstracts and provided feedback for improvements to annotation guidelines. M.G. provided ideas for several analyses. S.H.S. and M.G. oversaw the project progress and provided feedback on the design of study. All authors participated in the drafting and revision of the manuscript.

Abstract

Natural language processing (NLP) techniques can enhance our ability to interpret plant science literature. Many state-of-the-art algorithms for NLP tasks require high-quality labelled data in the target domain, in which entities like genes and proteins, as well as the relationships between entities, are labelled according to a set of annotation guidelines. While there exist such datasets for other domains, these resources need development in the plant sciences. Here, we present the Plant ScIenCe KnowLedgE Graph (PICKLE) corpus, a collection of 250 plant science abstracts annotated with entities and relations, along with its annotation guidelines. The annotation guidelines were refined by iterative rounds of overlapping annotations, in which inter-annotator agreement was leveraged to improve the guidelines. To demonstrate PICKLE's utility, we evaluated the performance of pretrained models from other domains and trained a new, PICKLE-based model for entity and relation extraction (RE). The PICKLE-trained models exhibit the second-highest in-domain entity performance of all models evaluated, as well as a RE performance that is on par with other models. Additionally, we found that computer science-domain models outperformed models trained on a biomedical corpus (GENIA) in entity extraction, which was unexpected given the intuition that biomedical literature is more similar to PICKLE than computer science. Upon further exploration, we established that the inclusion of new types on which the models were not trained substantially impacts performance. The PICKLE corpus is, therefore, an important contribution to training resources for entity and RE in the plant sciences.

Summary

In this chapter, I developed a high-quality labeled training dataset for NER and RE in the plant sciences. To the best of my knowledge, it is the first dataset of its kind specifically tailored to molecular plant biology, and consists of 250 documents labeled with biological entities and relationships between them. The development of a dataset for plant biology allowed us to evaluate the performance of existing NER and RE models in the plant sciences, as well as to train a joint NER/RE model specific to the plant sciences that improved information extraction on molecular plant science abstracts.

CHAPTER 3

DRYING TO CONNECT: UNIFYING THE RESEARCH LANDSCAPE OF DESICCATION TOLERANCE TO IDENTIFY TRENDS, GAPS, AND OPPORTUNITIES

The work in this chapter is presented in the pre-print:

Lotreck, S. G., Ghassemi, M., and VanBuren, R. T. (2024). Unifying the research landscape of desiccation tolerance to identify trends, gaps, and opportunities. *bioRxiv*

Author contributions:

R.V. and S.L. developed the initial project idea, and S.L. developed the idea for the conference recommendation algorithm. M.G. contributed ideas and discussion to the final implementation of the conference recommendation algorithm. S.L. implemented all analyses, made the raw versions of all figures and drafted the full text. R.V. provided input on figure organization and performed all final edits on the figures. R.V. and M.G. provided oversight on project progress and reviewed and edited the manuscript.

Abstract

Desiccation tolerance, or the ability to survive extreme dehydration, has evolved recurrently across the tree of life. While our understanding of the mechanisms underlying desiccation tolerance continues to expand, the compartmentalization of findings by study system impedes progress. Here, we analyzed 5,963 papers related to desiccation and examined model systems, research topics, citation networks, and disciplinary siloing over time. Our results show significant siloing, with plant science dominating the field, and relatively isolated clustering of plants, animal, microbial, and fungal systems. Topic modeling identified 46 distinct research topics, highlighting both commonalities and divergences across the knowledge of desiccation tolerance in different systems. We observed a rich diversity of model desiccation tolerant species within the community, contrasting the single species model for most biology research areas. To address citation gaps, we developed a rule-based algorithm to recommend new invitees to a niche conference, DesWorks, enhancing the integration of diverse research areas. The algorithm, which considers co-citation, co-authorship, research topics, and geographic data, successfully identified candidates with novel expertise that was unrepresented in previous conferences. Our findings underscore the importance of interdisciplinary collaboration in advancing desiccation tolerance research and provide a framework for using bibliometric tools to foster scientific integration.

Summary

Bob and I were particularly interested in performing analysis that would be of interest to other desiccation tolerance researchers by providing novel insights into the historical trends of citation and research topics in the field. I presented an initial version of this work and solicited community feedback at the DesWorks conference in January of 2024. Group debrief sessions during the conference inspired the design of an algorithm that could turn descriptive bibliometric analyses into a predictive tool that could provide actionable suggestions to improve research integration. To the best of my knowledge, the conference recommendation algorithm presented in this chapter is the first of its kind, and I have made the codebase with documentation publicly available so that it can be re-used and extended for other conferences.

CHAPTER 4

AN EVALUATION OF KNOWLEDGE GRAPH CONSTRUCTION AND AUTOMATED HYPOTHESIS GENERATION FOR WHOLE-PLANT DESICCATION TOLERANCE

Abstract

The proliferation of scientific information impedes the ability of scientists to keep up with new discoveries, especially in complex disciplines such as desiccation tolerance research. Desiccation tolerance, or the ability of an organism to revive from near-complete dehydration, is present across the kingdoms of life, but we lack an integrated understanding of the mechanisms of the phenotype. In this work, we aim to integrate information from across the drought and desiccation tolerance literature by constructing a large knowledge graph representing biological entities and their relationships. We evaluated several methods for knowledge graph construction, and found that neural network-based entity extraction, combined with co-occurrence-based relationships, provide the highest quality network. The resulting knowledge graph contains 334,327 biological entities and 1,288,387 relationships. Using two database-derived knowledge graphs and one other literature-derived graph, we provide preliminary evidence that literature abstracts may not be sufficiently information-dense to produce a high-quality connected network of biological entities, as database-derived networks had a consistently higher ratio of edges to nodes in our analysis. Using the co-occurrence network, we demonstrated that crop species are the most prevalent in the literature about drought and desiccation tolerance, and that while organism entities are the most common type of entity, that chemical compound entities are consistently the most well-connected across the literature. Finally, we applied knowledge graph embedding to build two kinds of static link prediction models to evaluate the possibilities for hypothesis generation from the co-occurrence network. We found that static link prediction, where the entire network is considered as a single snapshot, is insufficient to provide high-quality predicted hypotheses. We also explored a preliminary implementation of a temporal link prediction model, where the evolution over time of the network is considered during the link prediction task. While the static and temporal methods are not directly comparable to one another, we saw evidence that temporal link prediction may improve

upon the prediction capabilities of static link prediction. Our findings indicate future directions for improvement of hypothesis generation from knowledge graphs for biological literature.

Introduction

The scale and scope of biological knowledge are expanding exponentially, driven by both the increasing volume of published papers each year and the growing content within individual papers. This proliferation of information poses significant challenges for keeping up with new discoveries, even within narrowly defined fields, and it becomes even more daunting within large, multiscale, or complex disciplines. Knowledge integration across different disciplines is usually inadequate, leading to potentially important findings or connections between discoveries remaining unnoticed. This issue is particularly acute in the field of desiccation tolerance, a trait that enables organisms to withstand extreme dehydration. Desiccation tolerance is a widespread adaptation found across all kingdoms of life, prevalent in diverse organisms ranging from fungi and microbes to plants and animals. However, the knowledge spanning molecules to ecosystems remains fragmented and poorly synthesized. This work addresses this gap by attempting to leverage the extensive, yet disparate, body of literature to generate biological hypotheses concerning the genetic basis of desiccation tolerance in plants. This is achieved through the development and utilization of a knowledge graph to map out and connect information to identify underlying patterns and insights that might not be immediately apparent. By structuring data in this way, this research aims to enhance our understanding of desiccation tolerance, facilitating a more integrated approach to studying this crucial biological phenomenon.

A knowledge graph (KG) is a graph that contains data representing the real world, where the nodes are entities of interest, and edges are relations between them [Peng et al., 2023] . In biology, entities include proteins, genes, and organisms, and edges are relationships between entities, representing molecular interactions or regulations. The nodes and edges in a biological graph can be drawn from existing manually-curated databases, or they can be derived from unstructured text through natural language processing techniques [Nicholson and Greene, 2020]. Most graphs in the biomedical domain are constructed from existing databases [Nicholson and Greene, 2020],

and graphs in the plant science domain follow this trend. The large resource AgroLD is entirely derived from existing databases [Larmande and Todorov, 2021], and KnetMiner is primarily derived from databases, with some unknown level of supplementation from unstructured PubMed abstracts [Hassani-Pak et al., 2016, Hassani-Pak et al., 2021]. The more recent GenoPhenoEnvo graph [Thessen et al., 2023] is also constructed entirely from database sources. However, there are two examples of plant science KG constructed from literature-derived data: Comprehensive Knowledge Network uses a manual approach [Ramšak et al., 2018, Bleker et al., 2023], and PlantConnectome uses an automated approach [Fo et al., 2023]. To build a graph from the literature, we rely on information extraction methods, which include named entity recognition (NER) and relation extraction (RE). There are many approaches to NER and RE, including rule-based methods, and neural network-based methods. Rule-based methods that use syntactic (grammatical) rules, like OpenIE [Angeli et al., 2015], are domain-agnostic, while other rule-based approaches can incorporate domain-specific knowledge and be specific to a given subject area [Milošević and Thielemann, 2023]. Neural network methods tend to be domain specific, especially because they often assign entity and relation types to extracted objects, which are semantically relevant to a given domain, as seen in [Lotreck et al., 2023]. However, neural network methods can achieve higher performance on RE than domain-agnostic rule-based methods [Milošević and Thielemann, 2023]. To build a KG from literature, NER and RE are applied to a set of documents to obtain entities and relationship triples.

A "good" KG is information-rich, characterized by a detailed ontology that accurately represents real-world phenomena. According to Seo et al., "A good knowledge graph should have a fine-grained ontology structure that can precisely express information in the real world, and instances and triples should make full use of the ontology's classes and properties" [Seo et al., 2022]. For the biological sciences, a KG should contain genes, proteins, and organisms, with the relations between them indicating belonging or interaction and it should have as much real-world information as possible in each of those categories. Since our knowledge about biological life is far from complete , there will be plenty of missing information; however, that missing information should be the result of

true knowledge gaps, and not a failure to adequately capture or summarize the literature. It is therefore important to consider the data sources we use in constructing a KG. The proliferation of database-derived graphs as opposed to literature-derived KG could indicate either (1) that existing NER and RE tools are insufficient to extract information from text in the biological domain, but that sufficient information for high-quality graph construction is present in the literature; or (2) that there is insufficient information present in the literature to construct a high-quality graph.

The first part of the present work aims to determine which, if either, of the previous suppositions regarding the lack of literature-derived graphs is true. Here, we first constructed a large dataset (> 80,000 abstracts) of drought and desiccation tolerance literature, and examined four KG construction methods applied to this dataset. We then sought to determine, given the best possible KG from a literature source, how well we can generate novel hypotheses via link prediction. Link prediction is the act of predicting, based on the current structure of the graph, what information might be true, but is missing from the graph. Specifically, this takes the form of predicting the edges (or links) that are missing between entities in the graph [Rossi et al., 2021]. KG link prediction has traditionally been formulated as a static problem, where the entire graph is considered as a single snapshot, and edges are predicted on that snapshot; however, since KG reflect real-world information, which naturally evolves over time, treating the graph as a static item can lead to poor prediction performance [Cai et al., 2023]. Temporal link prediction (TLP) is the practice of predicting new connections between nodes at future timepoints for a given graph [Qin and Yeung, 2024], and can be used to improve link prediction on KG [Cai et al., 2023].

In general, graphs fall into two categories: homogeneous graphs, like social networks, where there is one type of edge and one type of node; and heterogeneous graphs, where there are multiple edge and node types. The power of KG is their ability to represent data from multiple sources with multiple kinds of relationships, which means they are an instance of a heterogeneous graph [Cai et al., 2018]. Unfortunately, while TLP is relatively well-developed for homogeneous graphs [Qin and Yeung, 2024], TLP for heterogeneous networks is limited to only a handful of methods, and does not have a comprehensive literature survey to describe the field as a whole. Meanwhile, static

link prediction is also well-developed for heterogeneous graphs. In this work, we first leveraged the existing computational resources for static link prediction to generate graph embeddings for a desiccation KG, and use both a simple Random Forest approach and an embedding-based ranked link prediction approach to predict new triples in the graph. We used the best static link prediction model to predict novel links between biological entities in our dataset, and perform a literature search to investigate the biology of a subset of the predicted links. Finally, we performed a brief survey of the literature on heterogeneous TLP, and implemented a method called STHN on our data to determine if TLP offers a performance advantage over static link prediction.

Results and Discussion

Characterizing the drought and desiccation tolerance literature

We built a combined dataset of drought and desiccation tolerance from Web of Science using two searches: "desiccation OR anhydrobiosis" and "(water deficit AND plants) OR (drought AND plants)". After post-processing the search results, the final dataset spans from 1985 to the present (**Figure 4.1A**), and contains mostly drought literature, with a small subset of the drought and desiccation literature overlapping with one another (**Figure 4.1B**).

Importantly, while our drought search on Web of Science specified that the papers should be about plant drought stress, the desiccation tolerance dataset includes papers from all kingdoms. We kept the non-plant papers in the desiccation tolerance literature in our combined dataset because there is already very limited information on desiccation tolerance, and we did not want to further restrict our data since we know that many mechanisms are shared across kingdoms. There is, however, an enormous amount of literature on plant drought tolerance, so we only included plant science papers in the drought portion of the dataset to keep the combined dataset to a computationally tractable size.

Defining a quality measure for a plant science knowledge graph

The goal of the present work is to build a knowledge graph of the desiccation and drought tolerance literature in order to make predictions about genes involved in the regulation of desiccation

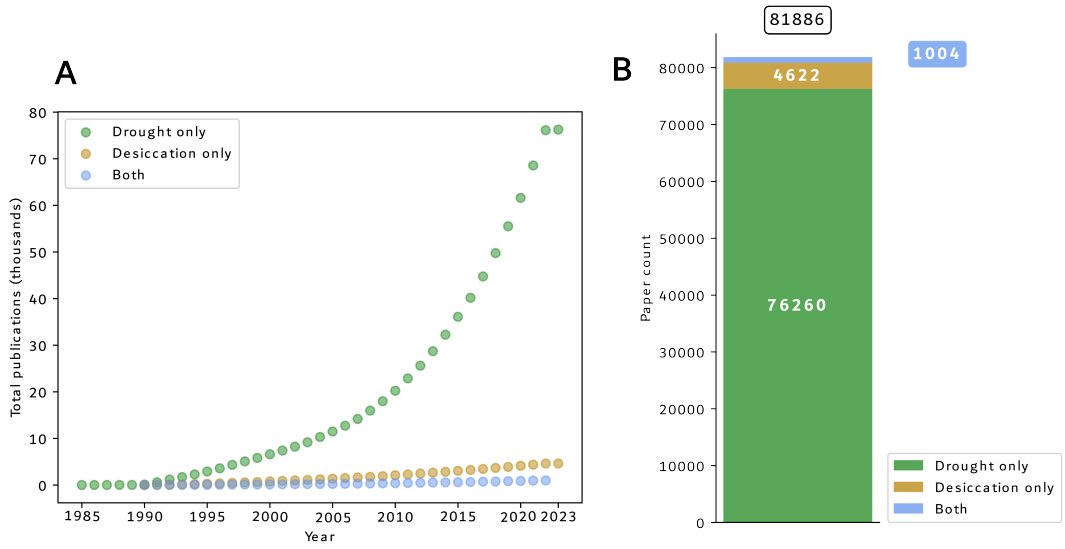


Figure 4.1 **Dataset statistics.** (A) Cumulative publications per year for drought, desiccation, and shared papers. (B) Number of papers in each category of the dataset.

tolerance. In biology, a knowledge graph is a network where the nodes are biological entities, such as genes, proteins, and organisms, and the edges are relationships between those entities. We aim to extract biological entities and their relations from scientific abstracts, which are an unstructured source of data, using named entity recognition (NER) and relation extraction (RE) methods.

Before we begin constructing a knowledge graph, we must define how we will evaluate the quality of the graph. KG quality evaluation is a non-trivial task, as different aspects of the KG are important in quality evaluation, depending on which downstream tasks the KG will be used for. Chen et al. calls this evaluation whether the KG is "fit for purpose" [Chen et al., 2019]. Differing requirements for KG quality in different scenarios means that evaluating a KG is not as simple as an accuracy or F1 metric like we could use for a classification algorithm. However, there exist many proposed metrics and frameworks for KG quality evaluation [Chen et al., 2019, Issa et al., 2021, Seo et al., 2022, Wang et al., 2021b]. In particular, we are concerned with the

quality of the KG related to the KG construction approaches that we use. Wang et al. specifically discusses quality control and evaluation during KG construction steps, breaking it down into three parts: (1) knowledge source selection, (2) knowledge extraction, and (3) knowledge fusion [Wang et al., 2021b]. In this work, we have selected scientific abstracts as our knowledge source. Wang et al. considers knowledge source selection principally from the perspective of credibility and relevance, including potential sources such as websites, crowdsourced information, and databases. We used a Web of Science search to choose abstracts for our dataset, which implies both credibility and relevance. In terms of knowledge extraction, we employ several NER and RE methods to build our graphs, and Wang et al. emphasizes the importance of limiting errors during the information extraction process [Wang et al., 2021b]. We can therefore consider the performance of our NER and RE methods to be a metric of the quality of the constructed KG.

However, traditional evaluations for NER and RE necessitate labor-intensive gold standard datasets labeled with entities and relations. We do not possess a labeled dataset for the domains of drought and desiccation tolerance, so we need to leverage the existing plant science dataset created in [Lotreck et al., 2023] (the PICKLE dataset). While we cannot directly determine if the entities and relations extracted from the drought + desiccation dataset are correct, we can create a proxy metric. Anecdotally, we notice that while NER seems to perform as expected across several of the methods we tried, exhibiting a relatively high estimated recall, barely any relations are extracted from any abstract with any method. Therefore, we will use the ratio of edges (relations) to nodes (entities) in the final extracted knowledge graph to determine if the NER and RE quality is in the general ballpark that we would expect for a dataset in the molecular plant sciences, using the PICKLE dataset to define our expectation for the ratio in a perfect NER/RE scenario. We will also make more direct comparisons of NER in the following sections to support the anecdotal observation that NER performs well, to further support using the relation:entity ratio as a measure of information extraction quality.

To support the validity of using PICKLE to generate a baseline expectation for an edge to node ratio for the drought + desiccation dataset, we first compared some basic dataset statistics, such

as the distributions of the number of sentences per abstract, the number of words per sentence, and word length for each dataset (**Figure 4.2A**). We see that both datasets exhibit nearly identical distributions, but that the drought + desiccation dataset, which is several orders of magnitude larger than the PICKLE dataset, has a very small number of outliers with larger values for each statistic. The statistical similarity of the two datasets, combined with the semantic similarity of the datasets (both molecular biology datasets with slightly different foci in terms of biological phenomenon), indicates that we can expect similar quantities of entities and relations to be extracted per document. In **Figure 4.2B**, we see the distribution of the edge to node ratio per abstract in the dataset. There are no abstracts with more relations than entities, and many documents have no relations, resulting in an overall edge to node ratio of 0.34. We will use both the per-abstract distribution and the overall ratio to perform a heuristic assessment of the graphs we build in the next section.

An important consideration when evaluating these networks is that scientific abstracts with sentence-level relation extraction may not be sufficient for constructing high-quality KG, even if NER and RE are performing perfectly to extract the available information in each abstract. After evaluating the approximate NER and RE performance of each graph construction algorithm in the following section, we will examine indicators of knowledge source incompleteness.

Knowledge graph construction methods struggle to identify semantic biological relations in text

We employed four graph construction approaches on our dataset: (1) DyGIE++, which is a joint NER/RE model [Wadden et al., 2019], (2) a co-occurrence approach using the entities derived from DyGIE++, (3) OpenIE, which is a rule-based method that uses syntactics to determine relations [Angeli et al., 2015], and (4) OntoGPT, which passes a predefined schema to GPT-3.5 for entity and relation extraction [Caufield et al., 2024]. The basic statistics of each resulting graph can be found in **Table 4.1**. Due to the proliferation of meaningless or unusable triples in the OpenIE results (see **Figure S4.1** for examples), we filtered out any triple whose entities did not appear in the DyGIE++ entities. Filtering brought the OpenIE results down from 323,233 entities and 644,175 relations to the values in **Table 4.1**.

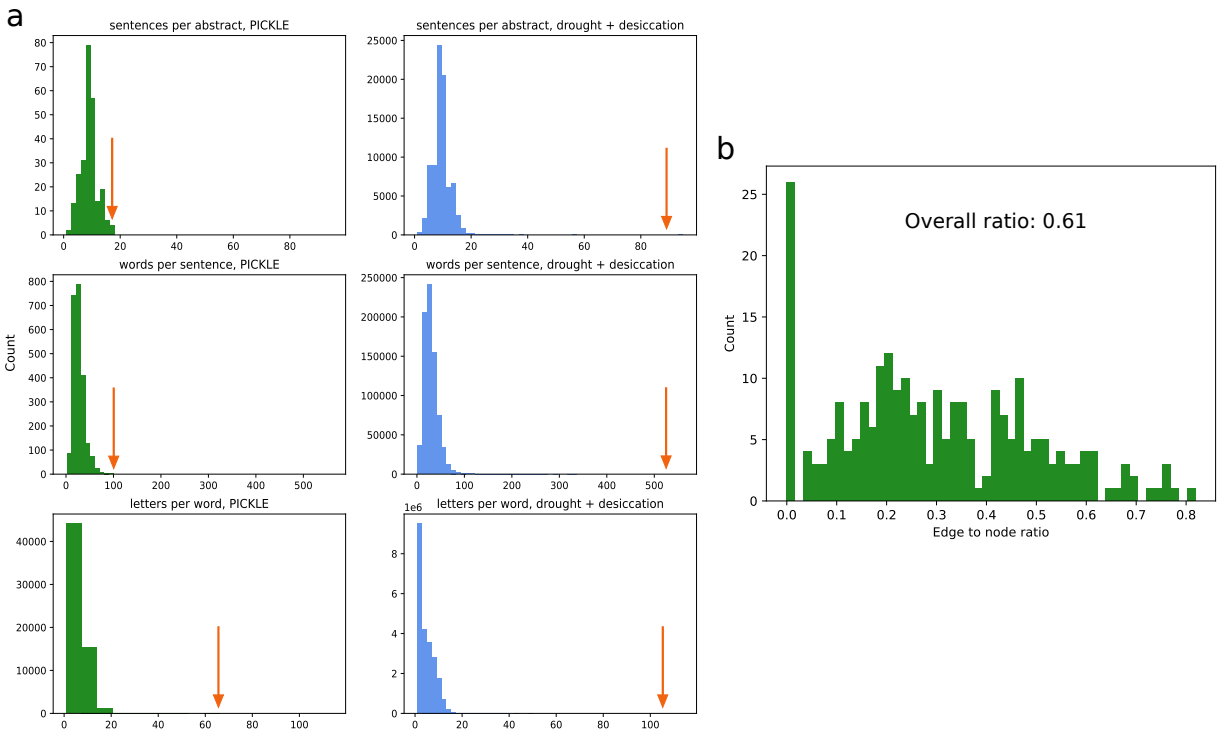


Figure 4.2 **Comparative dataset statistics and quality evaluation baseline.** (A) Histograms of basic dataset statistics for the PICKLE and drought/desiccation dataset. X-limits for each row are determined by the automatic x-limits for the drought/desiccation dataset, as it has larger outliers in each category. Default number of bins was used for PICKLE, and 5x the number of PICKLE bins was used for drought/desiccation in each row to allow a similar level of granularity for comparison. Orange arrows indicate the value of the maximum value in each plot. (B) Distribution of the edge to node ratio per abstract in the PICKLE dataset. The overall edge to node ratio is 0.34 for the dataset as a whole.

Method	# Nodes	# Edges	# Isolates	Median degree
DyGIE++	336,120	124,408	268,851	0
Co-occurrence	334,327	1,288,387	35,055	3
OpenIE	6,195	8,156	0	1
OntoGPT	12,488	3,023	9,981	0

Table 4.1 **Basic graph statistics.** Figures reported are after any cleaning performed on the raw constructed graphs.

Figure 4.3A shows the overall edge to node ratios for each of the graph construction methods. We see that DyGIE++ on its own, which attempts to extract semantic relationships, has a lower edge to node ratio than PICKLE. Given the statistical similarity between the two datasets, and the fact that PICKLE was designed specifically for use with DyGIE++, the lower edge to node ratio indicates that the DyGIE++ model is likely performing poorly on semantic relation extraction on the drought + desiccation dataset. In contrast, using sentence-level co-occurrence with the DyGIE++ derived entities yields a much higher ratio. This is expected, as co-occurrence cannot identify semantic relationships, and instead relies on the assumption that two entities that appear together in a sentence are related to one another. Without a gold standard, there is no way for us to quantify what proportion of co-occurrence relationships represent actual biological relationships. We can hypothesize, however, that since the ratio is substantially higher than that of PICKLE, there are likely many false positive relationships in the co-occurrence dataset. OpenIE is the only other construction method with an edge to node ratio greater than 1. However, OpenIE only extracts triples, and is not capable of directly extracting entities, which means that there are no isolate nodes, and a ratio of higher than 1 is guaranteed. OntoGPT displays an edge to node ratio relatively similar to that of the DyGIE++ method; however, it extracted an order of magnitude fewer entities than DyGIE++, which indicates that it is likely not suitable as a construction method. **Figure 4.3B** shows the distribution of edge to node ratios when calculated on a per-document basis, and there is a dramatic difference between the distribution of the per-document ratio of PICKLE and of all other methods, with all methods having a substantial right skew in their ratio distributions. The heavy skew of all methods indicates that relation extraction performance is particularly poor, as we would expect a more even distribution of ratios with more documents having non-zero ratios (i.e., having relations).

OntoGPT was a particularly promising method, as it grounds entities to databases in addition to using GPT-3.5. Schema grounding is intended to limit model hallucinations, and performance is drastically improved when using grounding [Caufield et al., 2024]. However, grounding is extremely slow, and even after optimizing grounding speed by using slimmed versions of databases,

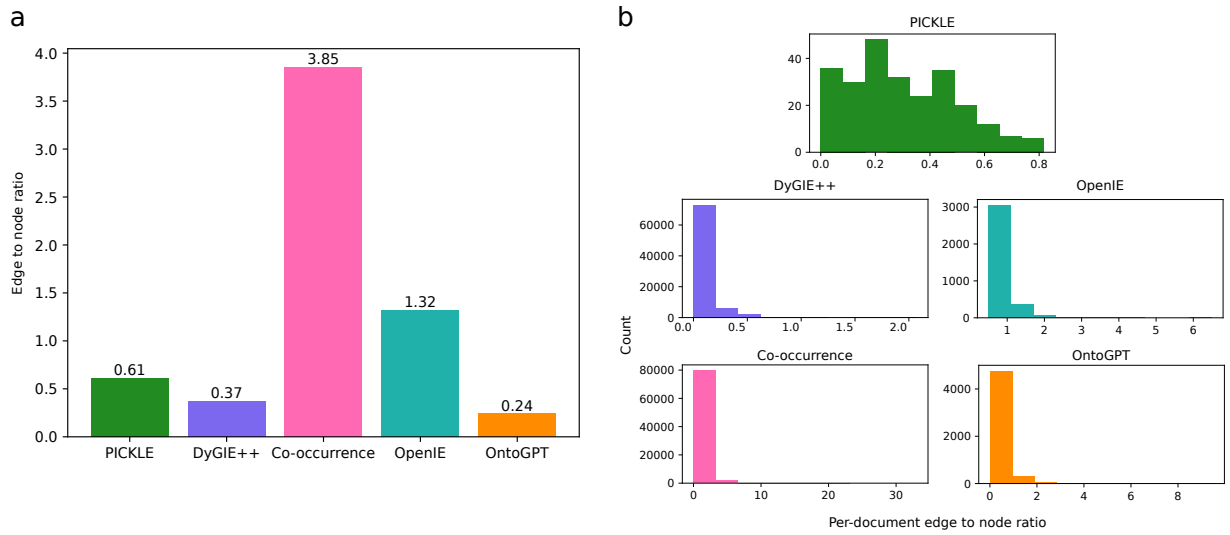


Figure 4.3 **Edge to node ratios for KG construction methods.** (A) Overall edge to node ratios for each construction method. (B) Distribution of edge to node ratios calculated on a per-document level for each construction approach. Note that there are only 5,237 documents in the OntoGPT graph, because of issues with computational complexity.

it would have taken 55 days to run OntoGPT on our whole dataset. We therefore only ran OntoGPT on the 5,237 document desiccation tolerance subset, using the slim NCBI Taxonomy to optimize computational performance. We found that schema grounding did not completely limit model hallucinations, especially when it came to relation extraction. The model hallucinated relations between non-existent entities like "NaN" and "Not provided". While hallucinated entities only made up 0.17% of the total extracted entities, relations that included one or more hallucinated non-entity comprised 48.34% of all extracted relations. An additional 5.34% of extracted relations were trivial relations between an entity and itself, and were also dropped. While entity extraction in general did not contain hallucinated entities, only 20.99% of entities were grounded back to one of the requested databases, with the remaining 79.01% simply receiving auto-generated unique identifiers that do not pertain to any database. None of our other methods included a grounding component, so even ~21% grounding is an advantage. That being said, by using TaxoNERD to ground just the Multicellular_organism DyGIE++ entities (see the next section for further details), we achieved 15.36% overall grounding, which indicates that OntoGPT does not achieve especially good performance over other methods for grounding entities externally to the KG construction

method.

Given the order of magnitude discrepancy in entities extracted by DyGIE++ and OntoGPT, we characterized the differences in NER between the two methods to get a sense of which is performing better. One important difference between the two methods is that while DyGIE++ extracts entities on a per-sentence basis, OntoGPT extracts them on a per-document basis, meaning that DyGIE++ can extract the same entity multiple times. To account for this, we resolved all entities with identical lowercase strings from each DyGIE++ document in this analysis, to avoid over-crediting extracted entity counts to DyGIE++. For each document in the desiccation tolerance subset, we quantified the proportion of the DyGIE++ and OntoGPT entities that were also identified by the other method ("shared", **Figure 4.4A**). We see that the distribution of the proportion of DyGIE++ entities is right-skewed, indicating that most documents have entities that were not identified by OntoGPT, while the OntoGPT distribution is left-skewed, indicating that almost all entities identified by OntoGPT were also identified by DyGIE++. When we look at abstracts randomly selected from the dataset (**Figure 4.4B**), we see that DyGIE++ identified many more entities than OntoGPT. One consideration to keep in mind is that the OntoGPT model was only tasked with extracting gene, protein, molecule, and organism entities, while DyGIE++ is capable of extracting some other types, like the Biochemical_process, Biochemical_pathway, or Plant_region. However, the extra types only account for a small portion of the DyGIE++-identified entities, and OntoGPT did not identify almost any entities of the types shared by both models. In the first abstract in **Figure 4.4B**, OntoGPT didn't identify any entities, and in the second, while it successfully identified two of the mosquito species, it hallucinated a third. *Aedes flavopictus* is a real species of mosquito, but it is not the same as *Aedes albopictus*, which is the species actually mentioned in the text.

After performing the above analysis, we decided that our best graph from these construction options is the DyGIE++-based co-occurrence graph. We immediately eliminated OpenIE on the basis of its proliferation of nonsensical/unusable triples, because when we filtered based on the relatively reliable entity set from DyGIE++, there were two orders of magnitude fewer entities and relations when compared to the DyGIE++ graph. The choice to eliminate OntoGPT as an option

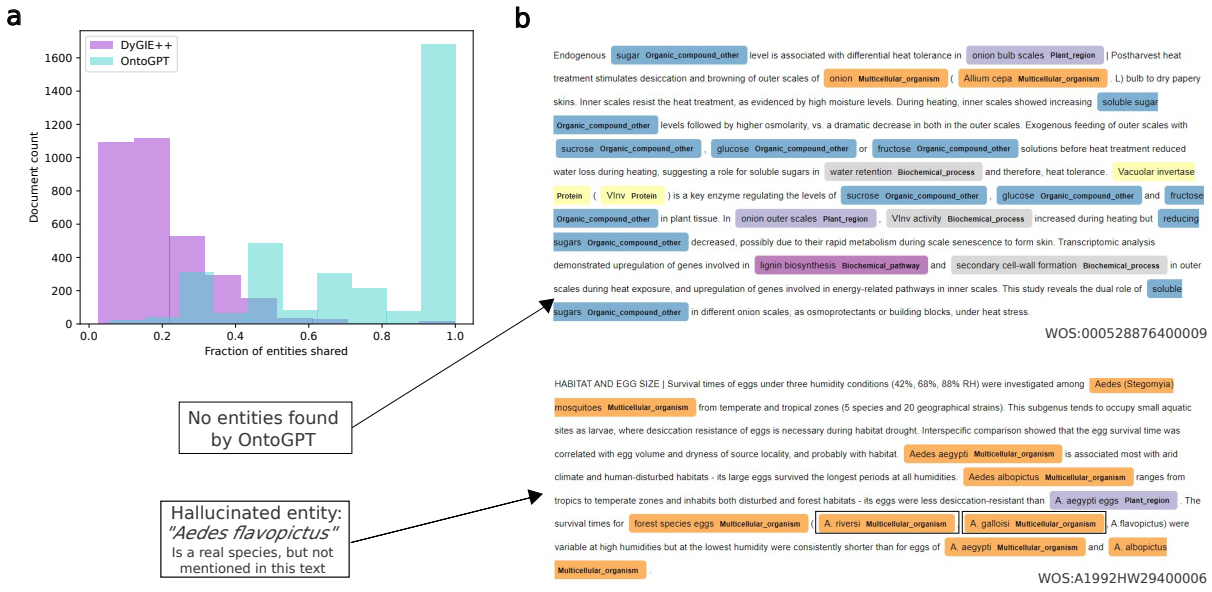


Figure 4.4 **Comparison of NER between DyGIE++ and OntoGPT.** (A) For each method, distribution of the fraction of entities in each document that are shared by the other method. (B) Example abstracts from the dataset with DyGIE++ entities annotated in colored boxes, where each color correpsonds to the entity type. OntoGPT-identified entities are outlined in black boxes.

was more complex, and involved both semantic and computational performance considerations. Firstly, the computational complexity of the grounding component of OntoGPT was prohibitive to running the method on our entire dataset. To confirm that grounding was necessary, we ran OntoGPT with a schema that contained no databases for grounding. While it did run extremely fast, most abstracts had no entities, and those that were extracted were nonsensical, which demonstrated that grounding is necessary. To make running OntoGPT on just the desiccation subset feasible, we had to substitute a slimmed version of the NCBI Taxonomy, which also anecdotally affected performance when we manually observed the output, both in terms of entities identified as well as their groundings. OntoGPT schemas can have prompts for each entity and relation type that are passed to GPT-3.5, and we provided prompts for all relation types, and for the Organism and Gene entity types. While we potentially could have further refined the prompts for entities and relations to tune performance, we chose not to move ahead with OntoGPT as our construction method. Even if good performance could be obtained, which seemed unlikely given our initial results, it would

have been prohibitively costly to run on the entire dataset.

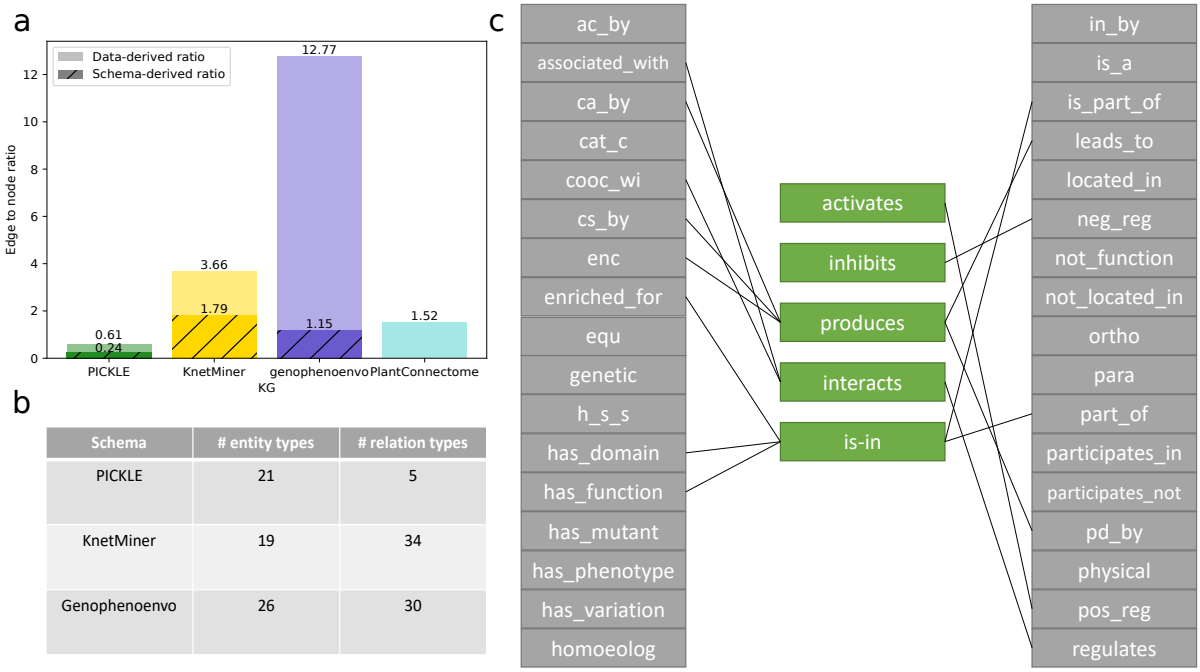
The elimination of OpenIE and OntoGPT as graph construction methods meant our choice was between the DyGIE++ and co-occurrence construction methods. We chose the co-occurrence method for two reasons. First, the DyGIE++ method on its own struggled to extract semantic relations from text, producing an edge to node ratio that was only slightly more than half of the ratio produced by the PICKLE gold standard. Secondly, co-occurrence has been used to excellent effect in many previous works, most famously in identifying a causal link between fish oil and the treatment of Raynaud's disease by Don Swanson in 1986 [Bekhuis, 2006, Swanson, 1986]. Additionally, while co-occurrence networks have the tendency to overestimate the presence of meaningful semantic relationships between entities, lowering a measure of specificity, it has been demonstrated that they have higher sensitivity in a biomedical use-case [Wang et al., 2021a]. High sensitivity indicates that a co-occurrence network likely contains a greater quantity of correct semantic relationships, even while it contains a larger volume of noisy links that don't reflect true semantics. Therefore, in the following sections, we will use the co-occurrence network in our analyses.

Scientific abstracts may not be a sufficient data source for a well-connected plant science knowledge graph

The difficulty of semantic relation extraction is clearly a limitation to using literature as a knowledge source in KG construction. However, it is important to consider the possibility that literature alone makes an insufficiently information-rich starting source for KG construction. To examine this possibility, we used two database-derived graphs, KnetMiner and GenoPhenoEnvo, and one literature-derived graph, PlantConnectome, to compute the edge to node ratio that we've been using as a proxy for information-richness thus far.

Figure 4.5A shows that both KnetMiner and GenoPhenoEnvo have higher edge to node ratios than either PICKLE or PlantConnectome does, which indicates that the database-derived graphs are more information-dense than a literature-derived graph. However, it is important to note that the schema for PICKLE, KnetMiner, and geonphenoenvo are not equivalent, meaning that they contain

different entity and relation types. It is possible that the difference in schema is responsible for the difference in connectivity density between the literature-derived graphs and the database-derived graphs, as the database-derived graphs have more entity and relation types. If this were the case, we would expect the actual edge to node ratio (the "data-derived" ratio) to scale with the ratio of relation types to entity types in the schema (the "schema-derived" ratio); however, this is not what we observe. While KnetMiner has the most relation types (**Figure 4.5B**) and as a result, the highest schema-derived ratio, GenoPhenoEnvo far outstrips KnetMiner in the data-derived ratio. Additionally, just because a schema has more type does not mean it can represent more information, as both entity and relation types can have varying levels of semantic granularity. For example, the term "regulates" can encompass both "upregulates" and "downregulates". **Figure 4.5C** demonstrates this concept for the KnetMiner and PICKLE relation schema. The five PICKLE relations map loosely to about 15 of the KnetMiner relation types, meaning that the PICKLE relation schema semantically covers almost half of the KnetMiner schema, despite only having a seventh of the relation types by number. Therefore, the drastically lower edge to node ratio of PICKLE is likely more related to the data source as opposed to the schema. In contrast to PICKLE, KnetMiner, and GenoPhenoEnvo, PlantConnectome uses GPT in a schema-free extraction approach, and the types in the resulting network are freehand phrases chosen arbitrarily by GPT. This results in a proliferation of unique "types", and relation types in particular are subject to rambling type descriptions, such as "had greater levels of resistance than" or "indicated variation in", which are reminiscent of the predicates extracted by the rule-based method OpenIE (**Figure S4.1**). Because there is no schema, we could not calculate a schema-derived edge to node ratio for PlantConnectome. However, the data-derived ratio for PlantConnectome is higher than that of PICKLE (**Figure 4.5A**). The prompts used in PlantConnectome's GPT-based approach allow the extraction of document-level relationships, which could potentially be responsible for the increased edge to node ratio. These data support the hypothesis that sentence-level relation extraction from the literature is an information-limiting condition, and that document-level extraction could potentially aid in better literature-derived graphs. However, even in a case where relations were extracted freehand from



Schema	# entity types	# relation types
PICKLE	21	5
KnetMiner	19	34
Genophenoenvo	26	30

Figure 4.5 **Literature-derived versus database-derived graphs.** (A) Edge to node ratios for each graph. Solid bars are calculated from all nodes and entities in the graph, while hatched bars are the ratio calculated from the number of relation and entity types in the schema. (B) The number of entity and relation types in each graph schema. (C) A loose mapping of the KnetMiner relation types (grey) to the PICKLE relation types (green). While an exact mapping is impossible to establish, we can loosely map one schema to the other, which shows that the PICKLE relation types have a wider semantic range than the KnetMiner types.

the entire abstract, unlimited by a schema, PlantConnectome still has a much lower edge to node ratio than the database-derived graphs. Without a performance metric for the extraction that created PlantConnectome, it's not possible to untangle whether this is due to data source or method performance. However, it seems likely that both method performance and data source are at play in the resulting lower ratio.

Without being able to compare a literature and database graph that were built on the same schema, we cannot decisively conclude whether literature is capable of building a sufficiently information-rich biological KG. However, given the indications that literature may not be sufficient for a high-quality biological KG, it is worth reflecting on why. In principle, literature contains all of the necessary information to build a densely-connected, information-rich KG, as the database sources used by graphs like KnetMiner or GenoPhenoEnvo are manually curated from the literature.

However, even in an ideal scenario where our information extraction methods were perfect, manual curators differ from most automated methods in two important ways: (1) manual curators have access to full text, while our methods above rely solely on abstracts, and (2) manual curators are naturally performing document-level information extraction, as opposed to the sentence-level relation extraction to which most current methods are limited. Intuitively, abstracts contain a summary of the most important points of a given paper, and should in theory be sufficiently information-rich. However, it is unlikely that a manual curator would only use abstracts to find information, as there is much more detailed information available in the full text of an article. Additionally, not all biological relationships are stated in single sentences, and it may take a relatively high level of reasoning over a whole paragraph or set of paragraphs in the full text of a paper to identify the relevant relationships. Manual curation is undeniably superior in these regards, but it cannot keep up with the flood of new publications. Therefore, research to more comprehensively identify the weaknesses of different literature data sources, as well as research on the best ways to balance the up-to-date nature of the literature with the more robust nature of databases for KG construction, is necessary.

Crop species dominate the drought tolerance research landscape

In an ideal world, we could analyze the properties of the constructed KG to gain insight on research trends over time. As we have outlined above some weaknesses inherent in our graph construction leading to a lower-quality graph, we must be cautious in the interpretation of graph properties; however, we can still gain valuable insights into research trends. In **Figure 4.6A**, the visualization of the entire graph shows that there is no sub-structure or neighborhoods in the graph, just one large grouping of nodes. This is consistent with the construction method of co-occurrence, as using sentence-level co-occurrence makes any entity nearly as likely to be connected to any other.

One of the weaknesses of the DyGIE++ method as trained on PICKLE is that there are no coreference capabilities, and we are unable to ground entity mentions back to database entries. However, we can partially address this by using external grounding methods, such as TaxoNERD

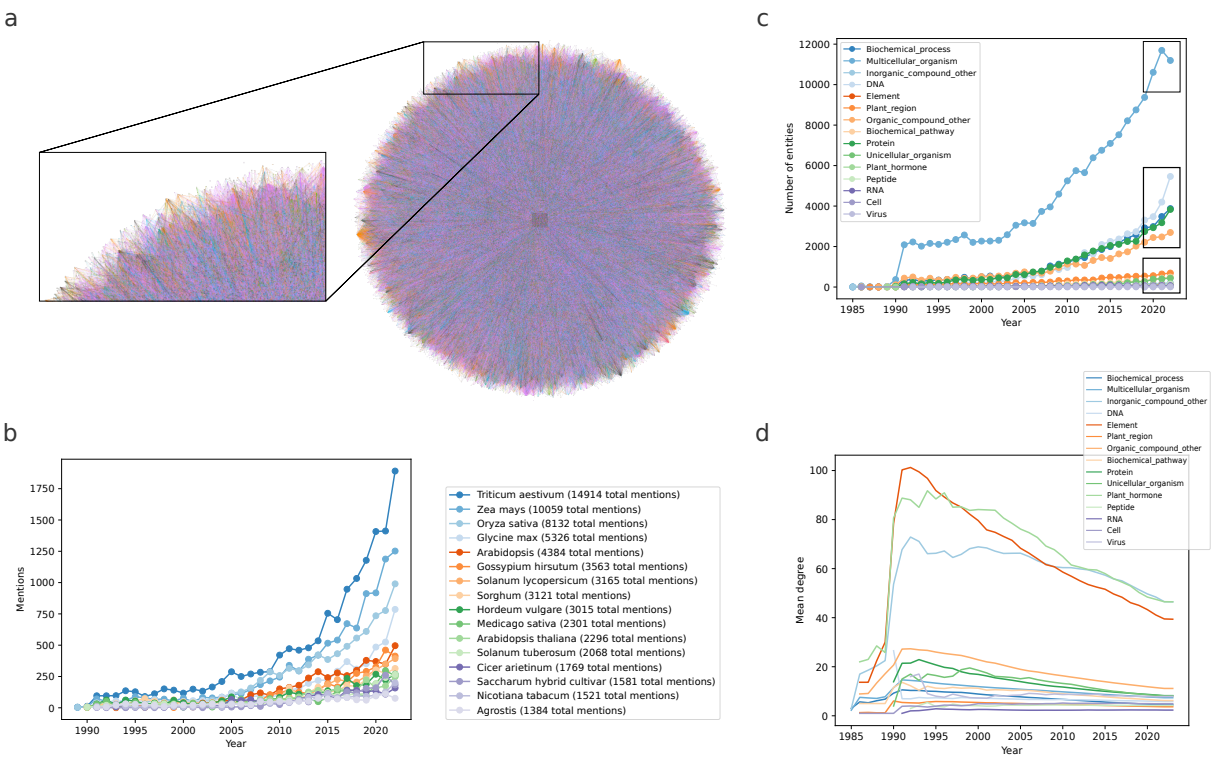


Figure 4.6 **Characterization of the drought-desiccation tolerance co-occurrence network.** (A) Overview of the entire network, with a zoomed-in detail. Nodes are colored by entity type, and edges are colored by their source node. (B) Grounded species prevalence in the graph over time; see Methods for details on the data pre-processing considerations for this analysis. (C) Prevalence of entity types over time. There are three groups in terms of growth trajectories, which are outlined in navy blue boxes. (D) Mean degree over time for all entity types.

[Le Guillarme and Thuiller, 2022]. To determine the extent by which we could ameliorate the lack of grounding with a single tool, we used TaxoNERD to ground all Multicellular_organism entities (see Methods for details). We found that 32.93% of Multicellular_organism nodes received a grounding; as nearly half of the nodes in the graph are made up of Multicellular_organism nodes, this means that 15.36% of the entire graph could be grounded with a single tool. We visualized the top 16 species and genus mentions in **Figure 4.6B**, and found that the most frequently mentioned species are crop species like wheat, maize, and rice. Combined, Arabidopsis and Arabidopsis thaliana are only mentioned about 6500 times, which, assuming (as we likely can do safely) that most Arabidopsis genus mentions refer to A. thaliana, means that research on drought tolerance of

wheat, maize and rice is more common than on A. thaliana as a model. The prevalence of crop species in this dataset is likely related to the agronomic importance of drought tolerance as a trait. In particular, mentions of the four major crops in the dataset are climbing at a faster rate than those of other species, indicating that research on drought tolerance in wheat, maize, rice and soy is increasing.

Another way to characterize the graph is to look at the growth of each type of entity over time. **Figure 4.6C** shows the growth of each entity type over time in the graph. We find three groupings of entity types by prevalence in the graph; Multicellular_organism entities are far and away the most prevalent, followed by a group that includes Biochemical_process, DNA, Protein, and Organic_-compound_other, and a third group containing the rest of the entity types. Interestingly, there was a leap in the number of Multicellular_organism entities in the early 1990's, followed by a period of relative stasis, followed by a second period of growth beginning around 2005. In contrast, other entity types have grown at a relatively constant rate. From 2010 onwards, as the middle group of entity types begins to take off, the growth of Multicellular_organism entities remains at a similar rate, indicating that while new organisms are still being added, more detailed investigation into the mechanisms of drought and desiccation tolerance in the organisms already in the graph is being undertaken. In contrast to which entities have the most nodes in the network, an entirely different set of entities are the most highly connected (**Figure 4.6D**), Element, Inorganic_compound_other, and Plant_hormone are the most highly connected node types in the network. Chemicals being the most well-connected makes sense given that many organisms and processes share connections to the same types of compounds. However, we also see that the mean degree of all node types decreases over time, indicating that there are more nodes in all types that have a very low degree. Because we don't have access to coreference resolution, this is likely due to a proliferation of unique string representations of semantically similar or identical entities as more and more new nodes are added to the graph, and not a reflection of any particular research trend.

Link prediction models are unable to generate biologically relevant hypotheses

To assess the possibilities of using link prediction (LP) methods on a literature-derived plant science KG, we applied the KG embedding model RESCAL on our co-occurrence network. RESCAL is a tensor factorization-based KG embedding method that yields embeddings for all nodes and edges in a given network [Nickel et al., 2011]. To provide a simple baseline forlink prediction performance, we trained a multi-class Random Forest (RF) model that took the embeddings of both nodes in a pair concatenated together as the feature vector, and predicted whether a given node pair should have a desiccation edge, drought edge, both desiccation and drought edge, or no edge. We tested two RESCAL loss functions and three negative sampling methods to optimize model performance (**Figure S4.2, S4.3** see Methods for details), but found that a random negative sampling strategy yielded the best performance (F1 = 0.30, AUROC = 0.64, **Figure 4.7A**). There was no meaningful difference between the RESCAL loss functions in the performance of the RF models, so we selected the RESCAL model trained with BCEWithLogitsLoss to be compatible with the RESCAL models used directly for prediction (**Figure 4.8**). There are several interesting aspects to the RF model's prediction capabilities. Notably, the RF model is unable to predict edges that only appear in the drought dataset, in fact achieving an AUROC score that is worse than random guessing, but is much better at predicting negative triples or triples that appear in both the drought and desiccation dataset (**Figure 4.7B**). The ability to predict negative samples is expected, as our training set contained as many negative instances as the total sum of positive instances (see Methods for justification); however, the model has clearly overcompensated and assigns the negative label to many instances that should be positive labels. This is an acceptable starting point for a model that is designed to generate testable hypotheses, as false positives are harmful in a scenario where a false positive means wasting resources and years of effort on predicting something that has no basis in reality. However, the model also similarly assigns instances to the both class with relative frequency, which is not explained by the presence of both in the training set, as there are an identical number of both, drought, and desiccation instances. Further investigation is needed to provide more comprehensive explanations of model behavior; however, the method as it stands is unsuitable for

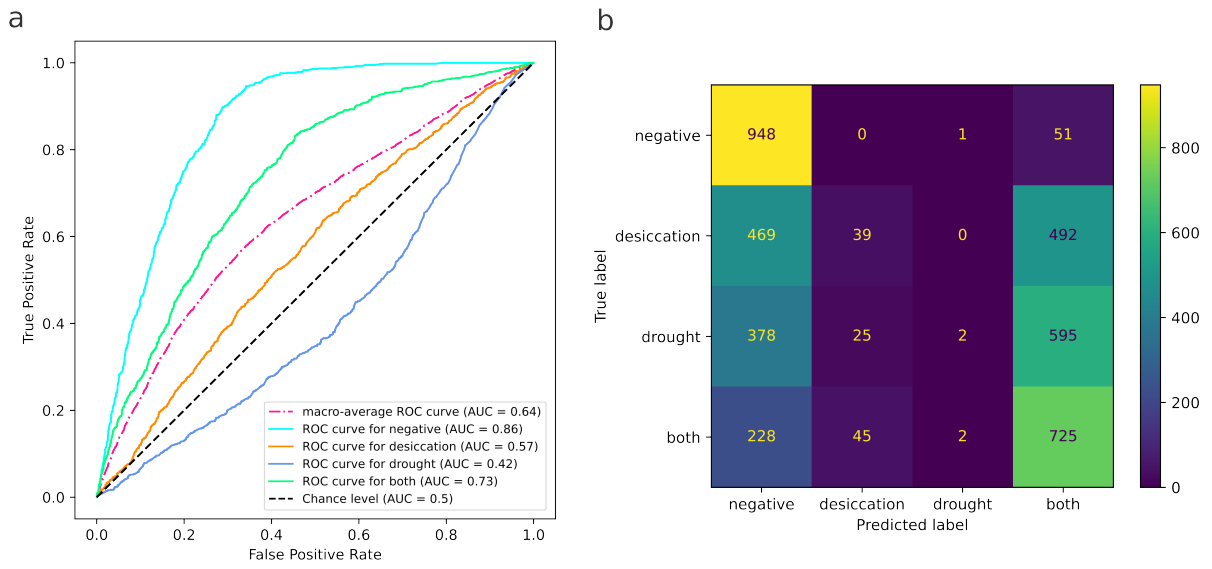


Figure 4.7 **Performance of Random Forest baseline models.** ROC curves for (A) the drought + desiccation model and (B) the GenoPhenoEnvo model. Note that in (B), classes 5 and 12 have true perfect performance, while classes 1 and 9 are just so high that they are rounded to 1.0. Confusion matrices for (C) the drought + desiccation model and (D) the GenoPhenoEnvo model. Note that for (D), the test set is imbalanced, so the color map doesn't visibly reflect the perfect performance of class 5, as it is very small.

hypothesis generation due to low performance.

Most KG embedding models are trained with a loss function that evaluates the model's capabilities for link prediction. Therefore, we wanted to see what the native prediction capabilities were for the RESCAL model we had trained. We employed two approaches to evaluating RESCAL's performance in link prediction. The first, to provide partial comparability with the RF model, was to ask the RESCAL model to generate a plausibility score for each of the 4,000 triples in the test set that we used for the RF model. We'll refer to this as the "predict triples" approach, after the function in PyKEEN used to perform this kind of prediction. KG embedding link prediction functions somewhat differently from the RF model we designed – rather than acting as a multiclass model that predicts both the presence/absence of an edge as well as its label, a KG embedding model provides a plausibility score that represents the model's confidence that the triple is true. The plausibility score can be leveraged with a threshold to generate a binary classifier by choosing a threshold plausibility score to make the cutoff between true and false for a given triple. When using

the BCEWithLogitsLoss, the model is optimized to score triples around a threshold of 0, so triples with a positive score are considered to be true, while triples with a negative score are considered to be false. Using 0 as a classifier threshold, the RESCAL model achieved an F1 score of 0.60, which is substantially better than the RF triple classification model scored. However, when we look at where positive an negative triples appear in the ranking, we see that negative triples appear at the top and bottom of the ranking, while positive triples tend to receive middling ranks (**Figure 4.8A**). Ideally, we would see that the bottom half of the ranking is predominantly negative triples, while the top is predominantly positive triples. To contextualize this finding, we can look at the distribution of triple scores for positive and negative triples (**Figure 4.8B**). Both the means and distributions of positive and negative triple scores are significantly different from one another (t-test, p-value = 0.004; KS test, p-value = 1.16e-43). In particular, the negative triples have a wider distribution of scores, with more triples at the high and low ends of the scoring range when compared to the positive triples. A wider distribution explains why negative triples appear both highly and lowly ranked, while positive triples tend to rank towards the middle. Because the BCEWithLogitsLoss is optimized specifically around a threshold of 0, we did not generate an ROC curve for this model; however, while the F1 score is substantially better than the F1 score of the best RF model (F1 = 0.30), looking at the rankings and distributions makes it clear that the predict triples approach is also insufficient to generate high quality assessments of link plausibility.

The second approach we took to using RESCAL for predictions is in line with what we would do to perform hypothesis generation for a new graph. We used the model to calculate plausibility scores for all possible triples in the drought + DT dataset, saving the top 100 scores. We'll refer to this as the "predict all" approach, after the function in PyKEEN used to perform this prediction. We manually investigated the links of the top 10 most plausible triples with a Web of Science search (**Table** 4.3). For each pair of terms, we performed an AND search to find papers where both terms co-occur. If no papers were returned and either of the two entities contained potentially superfluous terms that might confound the search, we simplified the search to increase the likelihood of obtaining results; for example, "wheat germ systems" was changed to "wheat germ". Final search

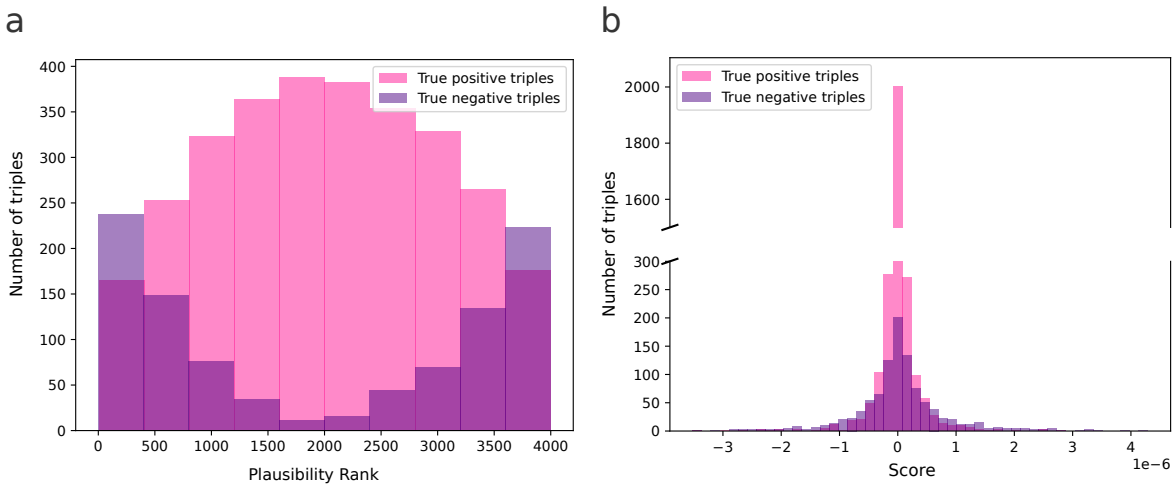


Figure 4.8 **RESCAL link prediction results.** (A) Rank distribution and (B) score distributions for the RESCAL model trained on the drought + DT dataset.

queries are detailed in **Table S4.1**. Generally, there are three categories of entity pairs. The first category are those relationships that hint at the presence of interesting biological relationships, but that lack the specificity of a good hypothesis. Triples in this category are ("*Crocus sativus L.*", "both", "tomatoes") and ("bZIP", "both", "deciduous forests"). The Web of Science results present a small subset of papers that hint at a mechanistic relationship between the head entities (*Crocus sativus L.* [saffron], bZIP) and some aspect of the tail entities. However, they lack the specificity to provide testable hypotheses; for example, how does saffron extract improve tomato resilience to stress on a mechanistic level? The second category of triples in the top 10 predictions are those that are trivial but true. For example, the triple ("beech-fir stand", "both", "deciduous forests") returns papers studying the ecology of forests; we know that tree stands occur in forests, and there is no implication of a further mechanistic relationship. Finally, there are those triples that are either irrelevant or incorrect, which return no papers when searched together, such as ("jasomic acid", "both", "*Amphibalanus amphitrite*"). Taken together, these results indicate that link prediction on KG as executed through an algorithm like RESCAL on a co-occurrence KG is currently insufficient to provide testable hypotheses at scale.

Importantly, our results demonstrate that blindly trusting performance metrics such as F1

Head entity	Edge type	Tail entity	# papers	Preliminary literature search reveals
wheat germ system	both	deciduous seasonal forests	1	Entomology report, "germ" in abstract is not wheat germ, appears separately from the term "wheat"
opuntia fragilis	both	deciduous seasonal forests	0	
pshsfa7a1_2595	both	mwsp	0	
salt stress-induced calcium signal	both	c. stelligera	3	None of the results mention calcium in the abstract
bzip23 transcription factor activity	both	deciduous seasonal forests	4	2 of the 4 papers identify bZIP transcription factors in trees (one deciduous species, one confier species), 2 papers are about transcriptional studies in trees but don't specifically mention bZIP
beech-fir stand	both	deciduous seasonal forests	23	Ecology studies in forests
t. fluminensis	both	deciduous seasonal forests	2	One paper mentioning ecological impact of T. fluminensis on forests, one that does not mention T. fluminensis
crocus sativus l	both	reds/breaker tomatoes	22	Search results are predominantly studies on the impacts of *Crocus sativus L.* (saffron) extracts on tomato growth)
drought-responsive and jasmonic acid biosynthesis genes	both	b. amphitrite	0	
mandarin water	both	deciduous seasonal forests	5	Two studies on mandarin ducks, three that don't mention the term "Mandarin"

Table 4.2 **PyKEEN top 10 prediction results.**

score or AUROC does not guarantee a model that performs well in context. Specifically, the RESCAL model achieved an F1 score of 0.60, which, when compared to the RF model's F1 score of 0.30, seems like a large improvement. However, we demonstrated that many negative triples are incorrectly classified as positive, indicating that the model is not useful in a practical context. Additionally, the AUROC scores of the RF models are misleadingly much higher than the corresponding F1 scores. Our findings highlight the importance of validating performance metrics with common-sense checks such as examining the most probable predictions to ensure that methods are providing practically valuable results.

As the performance of static link prediction on our dataset is exceedingly poor, we wanted to see what kind of prediction capabilities we can achieve using temporal link prediction models. We performed a brief literature review for heterogeneous TLP models, and identified 11 methods (**Table S4.2**). As a result of examining the available code bases for the identified methods and testing for functioning implementations, we selected STHN [Li et al., 2023], which embeds various aspects of the graph and summarizes information from a temporal interaction sequence to predict links. We ran STHN on the co-occurrence graph, and achieved an AUROC of 0.7685 after 74 epochs of training, but did not display a substantial increase in performance across training epochs (**Figure ??**). While the AUROC of STHN and our previous models are not directly comparable across model architectures, the higher AUROC of STHN indicates that incorporating temporal information into the link prediction task adds valuable information that can help improve predictions. Unfortunately, as the TLP field is relatively new and understudied, there are no robust codebases for any TLP method that we could identify. STHN was the most friendly, but while it was easy to run, it only outputs the AUROC scores, and does not save any predictions or allow re-use of the pre-trained model. Future work is required to delve deeper into the prediction capabilities of TLP for hypothesis generation on a plant science KG.

Conclusion

In this work, we have examined several KG construction methods and found that a combination of poor information extraction with low information richness in scientific abstracts results in

poor-quality KG. We provide preliminary evidence that literature abstracts may not be sufficient for high-quality KG construction when compared to database-derived sources. Using the best-constructed graph for the drought + DT dataset, we showed that crop species dominate the literature on drought and desiccation tolerance, and that while species names are the most common entities in the dataset, chemical compounds are the most well-connected entities. Finally, we explored the capabilities for hypothesis generation on a co-occurrence network derived from literature. We found that prediction capabilities for static link prediction based on a RESCAL KG embedding model are exceedingly poor, regardless of whether the built-in prediction capabilities or a downstream external method (Random Forest) were used. We additionally performed temporal link prediction on the co-occurrence KG, but due to limitations of the implementation are unable to further explore the results.

The largest limitation to high-quality KG construction in this study was relation extraction. Here, we showed that semantic relation extraction methods such as DyGIE++ and OntoGPT were insufficient to recover relations from natural language text. In this work, we chose to use a co-occurrence method to improve relation recall, as all other methods resulted in such sparsely connected KG that we were unable to implement downstream link prediction methods. However, there are clear limitations for co-occurrence as a method for constructing KG; principally among them is the over-representation of false positive triples in the resulting KG, which could be partially responsible for the poor performance of link prediction models. Further work on improving the quality of semantic relation extraction, either through the creation of more plant science-specific training datasets or through improved prompt engineering of large language models will likely result in high-quality KG.

Another important future direction of this work is to explore the potential benefits of using full text documents rather than abstracts. In particular, quantifying the richness of biological relationships in full text versus abstracts will be important to determining the optimal data sources for KG construction. Potentially, document-level relation extraction, made possible by tools like GPT, will also benefit information retrieval from literature, as relations may not be stated in single

sentences. While the literature is the most current and complete source of information, leveraging the data quality of manually-curated databases by integrating literature-derived information with database information through entity and relation grounding is also an important area for future work.

Finally, a deeper investigation into the performance of KG embedding models and link prediction on biological datasets is called for, as their performance here was exceedingly poor. Biological KG often have different underlying properties than the ideally-distributed graphs on which KG embedding models are often evaluated, and this may impact the ability of KG embedding models to effectively embed biological networks. The performance of KG embedding models is also likely tied to the performance of upstream relation extraction methods used to construct the KG on which prediction is performed. Specific to the analyses presented in this work, the use of co-occurrence, which likely result in a large number of false positive triples, could exert a confusing effect on the KG embedding model, as the model will generate embeddings that conflate the characteristics of true negative triples for those of positive ones. Additionally, biological KG exhibit a large skew in the distribution of degree, often approximating scale-free behavior, which can differ greatly from the datasets on which embedding models are evaluated in their original publications. Quantification of the impact of network topologies and of upstream relation extraction methods on KG embedding algorithms will be a valuable next step in determining whether, and how well, KG embedding models can perform in the link prediction task. In this work, adding a temporal element to link prediction seems to improve prediction capabilities on a preliminary basis; however, the same pitfalls that static link prediction methods experience on biological datasets could also have an impact on the ability of temporal models to produce high quality predictions.

Acknowledgements

We thank Max Berrendorf from the PyKEEN development team for his invaluable help in understanding the implementation of RESCAL models, particularly with hyperparameter selection. We also thank Harry Caufield of OntoGPT for his help in implementing and optimizing the schema for OntoGPT.

Methods

Dataset construction

We used the methods from the previous chapter to obtain and pre-process a dataset on drought, and combined it with the desiccation tolerance dataset from the previous chapter to create the dataset used here. We used the query "(TS=(water deficit) AND TS=(plants)) OR (TS=(drought) AND TS=(plants))" on Web of Science to obtain the drought dataset. We downloaded the first 100,000 results (of a total of 134,510 query results). In the end, only 99,598 entries were downloaded from Web of Science, as some of the Fast 5000 results were incomplete. We don't know why this is; however, since the dataset is still substantial even with $\tilde{4}00$ missing results, we chose to move ahead. 9,024 were dropped because they were outside our version of the XML dataset, and a total of 88,433 were recovered. When combined with our previously constructed desiccation dataset (5,963 documents), we obtained a total of 93,348 documents (Note that the discrepancy in addition is due to there being some documents in common between the two datasets). We then extracted abstracts to text documents, which resulted in the loss of an additional 11,169 documents because their XML entries did not contain abstracts. This gave us a final dataset of 81,886 documents; 76,260 drought abstracts, 4,622 desiccation abstracts, and 1,004 abstracts that appeared in the searches for both drought and desiccation.

Knowledge graph quality measure determination

To compare the two datasets for statistical similarity, we calculated the number of sentences per abstract, the number of words per sentence, and word length for each dataset, and plotted their distributions. We built a networkx graph for the PICKLE dataset, which removes duplicate nodes and edges. We calculated the ratio of edges to nodes for the overall PICKLE dataset based on the networkx graph, and then calculated the edge to node ratio on a per-document basis for each method. The per-document calculation allows repeated entities and nodes across the whole dataset, as the same entities can be extracted across multiple documents. We also allowed repeated entities within the same document, as sentence-level relation extraction methods like DyGIE++ rely on every instance of each entity being present.

Knowledge graph construction

We tested four avenues for KG construction from the combined drought and desiccation dataset (**Table 4.1**). The first two approaches rely on the DyGIE++ architecture, which is a joint entity and relation extraction model based on the idea that the properties of entities contribute information to the process of relation extraction and vice versa (see [Wadden et al., 2019] for architecture details). We used a DyGIE++ model that we had previously trained on the PICKLE dataset and applied it to the entire drought + desiccation dataset [Lotreck et al., 2023]. In the first construction approach, we used the output of DyGIE++ as-is, with no modifications. In the second approach, we kept the entities extracted from DyGIE++, but derived relations from sentence-level co-occurrence; if two entities appeared together in a sentence, we put an undirected relation between them in the resulting graph. We kept track of how many times each entity and relation appeared in the dataset, recording the total number of times each appeared, as well as their first date of appearance, and whether or not the relations were derived from a drought article, a desiccation article, or both. Our third construction approach was OntoGPT, a GPT-3.5-based approach to extract entities and relations to a predefined schema of entity and relation types (see [Caufield et al., 2024] for implementation details). OntoGPT uses entity grounding to databases specified in each schema to prevent GPT-derived hallucinations and to improve the recall of extraction. We built a schema to extract genes, proteins, molecules, and organisms, as well as the relationships between each of those types. To improve the likelihood of good extractions, we provided prompts for each relation type. Unfortunately, in the current OntoGPT implementation, grounding is extremely computationally complex and does not scale to larger datasets or larger schema databases. While it is possible to substitute slim databases or use no databases whatsoever, this severely impacts the quality of the extraction (see [Appendix] for a characterization of the information extraction and computational performance impacts of the various options), and still does not result in enough of a speedup to make implementation on a dataset larger than a few thousand documents practical. As a result of our analysis, we applied OntoGPT with a schema using the slim version of NCBI Taxonomy on our desiccation dataset only, as applying even the slimmed schema to the whole dataset was

prohibitively costly, and our initial performance evaluations did not provide sufficient justification for investing resources into further application. Upon manual inspection of the subset results, the OntoGPT graph contained a large number of hallucinated relations based on hallucinated entities such as "NaN" and "Not provided". We trawled the dataset for such entities and removed them, as well as any relations that depended on them. Finally, as a common-sense baseline, we applied the rule-based approach OpenIE, which extracts triples from text in a domain-agnostic manner using syntactic information (see [Angeli et al., 2015] for implementation details). As OpenIE has no knowledge of domain-specific considerations for writing style, it opts for a high-recall approach by extracting every possible triple, resulting in a large proportion of extracted triples being nonsensical or unusable (see **Figure S4.1**for examples). To combat this issue, we decided to keep only triples whose entities matched DyGIE++-extracted entities, as we had a high degree of confidence in our DyGIE++ results based on the performance evaluations in plant science presented in [Lotreck et al., 2023], as well as manual examination of a small subset of our results on the drought + desiccation dataset.

We calculated the whole-dataset and per-document edge to node ratios in the same way that we did for PICKLE; whole-dataset ratio was calculated using the networkx graph where duplicate entities and relations are resolved, and calculated the per-document ratios allowing duplicate entities and relations. For documents that have 0 nodes (which does not occur in PICKLE but can occur when an automated method is applied), which would result in a ZeroDivisionError on computation of the ratio, we substituted a 0 for the ratio to represent that no information was extracted.

To compare the NER capabilities of DyGIE++ and OntoGPT, we mapped the document ID's to the randomly-generated OntoGPT document ID's to pair up the entities extracted from each document in the desiccation tolerance subset. For each document of DyGIE++ entities, any lowercase entity strings identical to one another were resolved into a single entity before calculating the proportion shared. For each of DyGIE++ and OntoGPT, we calculated on a per-document basis the proportion of their entities that appeared in the intersection of the OntoGPT and DyGIE++ entities for each document, and plotted the distributions of those values for each method. We

Construction Method	NER approach	RE approach	Refinements
DyGIE++	Joint neural method	Joint neural method	None
DyGIE++ co-occurrence	Joint neural method	If two entities appear in a sentence together, a relation is placed between them	None
OntoGPT	GPT-3.5 extraction to predefined schema	GPT-3.5 extraction to predefined schema	Prompts added to relations in the schema, only ran on the desiccation tolerance subset
OpenIE	Rule-based, domain/schema agnostic	Rule-based, domain/schema agnostic	Filter initial output to only keep entities (and correspondingly their relations) that are included in the DyGIE++-extracted entities

Table 4.3 **Summary of KG construction methods.**

randomly selected 5 abstracts from the desiccation tolerance subset to visualize DyGIE++ and OntoGPT entities, and manually selected two to appear in the figure.

Comparison of graph connectivity

We selected two predominantly database-derived graphs for comparison to our networks: KnetMiner [Hassani-Pak et al., 2021], and GenoPhenoEnvo [Thessen et al., 2023].

To obtain a relevant edge to node ratio for KnetMiner, we used the KnetMiner Neo4j browser (http://knetminer-wheat.cyverseuk.org:7474/) for the Poaceae network, which contains wheat and Arabidopsis. We used sample Neo4j Cypher commands provided in the browser to get the entity and relation types present in the network. The KnetMiner network differs from other networks examined here because it also includes the data sources as nodes with relationships in the network. To avoid artificially altering the computed edge to node ratio, we sought to remove any entity and relation types that dealt with data sources, as opposed to biological entities or concepts. While we

were unable to locate documentation explaining each entity and relation type, most names were semantically sensical, and so we manually created a subset of both types that we believed to be biological in nature. To confirm that the less explainable entity types were in fact biological, we used the Cyhper command "MATCH(n) WHERE n:<entity type> RETURN n LIMIT 5" to examine the names and properties of the first five entities of each ambiguous type and determine whether the type was relevant. While this resulted in a relatively high-confidence list of entity types, the semantics of the relations were much less clear. Our goal was to include only relation types that are restricted to connecting entity types that were in the identified list of relevant entity types. While the database schema should have provided the necessary information, there was no information about what entity types were valid subjects/objects for the various relations. We therefore constructed the following Cypher command to get the entity and relation types for all relations in our proposed relation list: "MATCH (n1)-[r:<relation type>]-(n2) RETURN labels(n1), TYPE(r), labels(n2)". While it is possible to run this for all relation types at once with an "or" operator, the network is large enough that running one combined command crashed the web server for the browser, so we ran this command for each relation type in our proposed list separately. We then asserted that all entity types in all triples for the given predicate were in our list of biological entity types, and kept only relations for which this was true. The one exception was for the "part_of" relation, two of its relations contained CoExpStudy (co-expression study) entities; however, there are only two CoExpStudy entities in the entire graph, so we kept "part_of", as the entities it connects are predominantly biological. We then used the following two queries to count the number of entities and relations across all the biological types:

MATCH (n) WHERE n:Gene OR n:ProtDomain OR n:Path OR n:CelComp OR n:BioProc OR n:MolFunc OR n:EC OR n:Comp OR n:Protein OR n:Protcmplx OR n:Enzyme OR n:Reaction OR n:CoExpCluster OR n:SNP OR n:Transport OR n:Phenotype OR n:PlantOntologyTerm OR n:SNPEffect OR n:Trait RETURN count(*)

MATCH ()-[r:cs_by | in_by | participates_in | enriched_for | has_phenotype | ortho

| is_a | not_located_in | ca_by | pd_by | enc | leads_to | homoeolog | located_in | has_-function | ac_by | physical | has_variation | pos_reg | participates_not | has_domain | associated_with | is_part_of | h_s_s | para | neg_reg | cat_c | equ | regulates | has_mutant | genetic | cooc_wi | not_function | part_of]->() RETURN count(*)

We then used the two resulting values to calculate the edge to node ratio. The GenoPhenoEnvo graph is available for direct download as two dataframes, a nodelist and an edgelist. We downloaded these and used the Python package networkx to create a graph object for analysis. All types in GenoPhenoEnvo are biological, so we didn't perform any further pre-processing. For all graphs, we also used the number of entity and relation types to calculate a "schema-derived" (as opposed to "data-derived") edge to node ratio.

Graph characterization

We visualized the co-occurrence network in Gephi using the OpenOrd visualization algorithm, coloring nodes by their entity type and edges by their source node.

As the PICKLE dataset doesn't allow us to use the coreference option in DyGIE++ (which attempts to improve predictions by mapping different mentions of the same real world object to a single entity), we implemented a partial coreference resolution approach by grounding predicted entities in the Multicellular_organism class back to NCBI Taxonomy using the TaxoNERD model [Le Guillarme and Thuiller, 2022]. Rather than using the full text of the originating abstracts and allowing TaxoNERD to also perform the NER step as it would in its full pipeline, we used the DyGIE++-derived entities in isolation, applying the TaxoNERD entity linker by itself. We combined entities into spaCy documents up to the maximum allowed number of characters, specifying entity span boundaries, and then applying the TaxoNERD linker. Using this approach, 32.93% of Multicellular_entities received a grounding. Once grounded, we mapped entity names to their Taxonomy groundings, and summed the number of entities in each year that corresponded to each Taxonomy grounding to obtain growth trajectories over time for the top 20 most frequently mentioned. For this analysis, we ignored any entity that was not grounded. We examined a subset of the original entity names for groundings in the top twenty that seemed suspicious, and

identified several weaknesses with the TaxoNERD groundings. First, any entity that contained the word "transgenic" was mapped to Mus musculus, or the house mouse, likely because "transgenic mice" is a common entity. Many entities containing the phrase "___ plants", like "rice plants" or "olive plants", were mapped back to Embryophyta due to the presence of the word "plants", and phrases containing "grapevine" were mapped as "Grapevine virus A". Additionally, the DyGIE++ model has the bad habit of identifying country names as Multicellular_organism entities, so China appeared in our top twenty. We therefore removed these entries from the top 20, leaving 16 top species.

To examine the growth of each entity type category all the time, we used the full graph for all types (without grounding for Multicellular_organism entities). For each year in the dataset, we summed the number of entities with that year as their first mention for each entity category. We removed 2023 from the final visualization, because it is a partial year and therefore brings all entity type values near 0. To examine degree over time for each entity type, we sliced the graph at each year, removing all nodes (and as a result, edges) past the cut year, and calculated the degree for all nodes present in the graph.

Link prediction problem setup

In order to predict hypotheses, we need to develop a framework for how to use the KG in a prediction setup. Our co-occurrence graph contains biological entities, with undirected relations that have three possible labels: desiccation, meaning the link was derived from a paper in the desiccation tolerance portion of the dataset, drought, meaning it was derived from the drought portion of the dataset, and both, which means the paper that provided the relation is found in both portions of the dataset. At a high level, our goal is to predict what node pairs should have a desiccation or both designation. The theoretical grounding for this problem setup is that we want to leverage the much greater quantity of literature available in drought tolerance research to determine the genetic basis of desiccation tolerance. Potentially, the genetic elements identified as important in drought tolerance could also have a role in desiccation tolerance, as drought tolerance precedes desiccation tolerance when a plant begins to dry down. Therefore, we need a model that can predict

new edges of varying types on an undirected graph. We are interested in both static and temporal predictions: in the static case, we want to predict node pairs that should have desiccation edges based on a static snapshot of the graph over all time, while in the temporal case, we want to predict which node pairs should have an edge at the next time point based on the evolution of the graph over time.

Static link prediction

We tested two approaches to static link prediction on the DyGIE++ co-occurrence network. Both approaches are based on the KG embedding model RESCAL, as implemented in the PyKEEN package [Ali et al., 2021, Nickel et al., 2011] . We first wanted to see if we could design a model that correctly predicts the type of (or that there should not be a) relation between a given pair of nodes. As a simple baseline, we trained a Random Forest (RF) classifier, using the node embeddings derived from RESCAL as features. We tested two versions of the RESCAL model for generating the embeddings for RF features: one using the default MarginRankingLoss with default entity and relation initializers, and one using BCEWithLogitsLoss and the "normal" entity and relation initializers (4.4). We split the data into train/validation/test sets with the ratio 0.8/0.1/0.1 using the random seed 1234, and the same training/validation/testing splits were used for each RESCAL model. In addition to evaluating the impacts of the RESCAL loss function on the downstream RF model, we also tested three negative sampling strategies for the RF method: random, corrupted tail, and embedding-based.

Random: The random sampling approach aims to choose a random subset of the possible combinations of head and tail entities in the dataset. In practice, random selection from all possible combinations is computationally intractable for a dataset of this size; even just calculating the number of possible combinations causes an OverflowError. Therefore, we implemented a computational shortcut that approximates true random sampling of combinations. We randomly sample head and tail entities separately, and then pair the corresponding indices of the two lists together to create pairs, removing any pairs that either appear in the positive set, or that have identical head and tail entities.

Corrupted Tail: The corrupted tail sampling method is modeled after the default sampling method implemented in PyKEEN. In our implementation, each negative instance is created by taking one of the positive instances and replacing the tail entity with a randomly sampled other entity to create a triple that is not found in the positive set.

Embedding: In a KG, there is an enormous number of possible negative triples that far outweighs the number of positive triples, as any two nodes in the network not already connected by an edge can form a negative triple. However, not all negatives are created equal; it is more difficult for a model to identify a negative triple that is semantically plausible than one that is clearly false. Therefore, random sampling to generate the negative instances for the training set is likely to result in a model that is not able to successfully distinguish between positive and negative samples in a realistic case where the negatives are not obviously incorrect. To ameliorate this, we tried a corruption approach, where negative triples are generated by taking a positive triple and randomly replacing the tail entity, and an embedding-based approach. The idea of the embedding-based approach is to generate negative triples that are semantically plausible, to force the RF model to learn to distinguish between all kinds of negative triples and true positive triples. We modeled our embedding-based negative sampling on the method presented in [Islam et al., 2021]. For each positive triple in the dataset, we randomly sample 50 possible new tails, and eliminate any that would make a true positive triple. For each new tail, we use the embeddings from the RESCAL model to calculate the Euclidean distance between the original and all new possible tails. We then calculate a softmax probability on the Euclidean distances, which generates a score that is higher for tails that are closer to the original triple (have smaller distance scores). Following [Islam et al., 2021], rather than taking the highest-scoring triple directly, which can lead to accidentally sampling false negative triples (since the KG is known to be incomplete, negative triples may actually be true, unrecorded triples), we sample randomly from the top 5 highest scoring triples to choose the new tail. If the resulting triples is already in the negative set, we sample again until we obtain a triple that has not already been chosen.

For each upstream RESCAL model (MarginRankingLoss vs. BCEWithLogitsLoss), we trained

three RF models. Each model's training set had the same positive instances, but each model used one of the three negative sampling methods to generate the training instances in the negative class. We sampled the training instances out of the RESCAL training set, and the test instances out of the RESCAL testing set. All models were tested on the same test set, which was generated using the random negative sampling method and contained 1,000 instances of each positive class, and 1,000 negative instances. The training set for each RF model contained 2,000 instances of each positive class, and 2,000 * (number of positive classes) number of negative instances. While in principle a balanced training set would contain 2,000 negative instances if negative is considered a uniform class, there are negative triples that correspond to one class or another. For example, if a positive triple is (Barak Obama, is, Democrat), the negative triple (Barak Obama, is Republican) is a negative triple with similar semantics, while the negative triple (Photosynthesis, produces, Expo markers) belongs to an entirely different semantic grouping. Additionally, we would prefer our model to predict more false negatives than false positives, because in principle, any hypothesis predicted by this model would require potentially years of labor on the part of an experimental scientist, so being cautious in how we treat the negative class is prudent. In total, we trained and evaluated six models for the drought + desiccation dataset. For each RF model, we used sklearn and performed a random search hyperparameter optimization with the following parameter distribution settings: "n_estimators": randint(100, 500), "max_depth": randint(1,50), "criterion": ["gini", "entropy", "log_loss"]. We used the sklearn functions f1_score (average = "macro"), auc_roc_score (average = "macro" and multi_class = "ovo") and confusion_matrix to evaluate the output.

The second prediction approach we used was the built-in prediction functionality of PyKEEN's RESCAL, which calculates a score for the probability of a given triple being true. The PyKEEN implementation provides three basic prediction functionalities, which all rely on the calculation of a plausibility score: (1) calculating the score for every possible triple ("predict all"); this is not recommended by the developers as it necessitates calculating a score for every possible triple and is therefore computationally intensive; (2) calculating scores for a specific list of triples ("predict triples"), and (3) given a head entity and a relation type, returning an ordered list by plausibility

Parameter	Margin Ranking Loss Model	BCE Loss Model
stopper	"early"	"early"
model	"RESCAL"	"RESCAL"
model_kwargs		dict(entity_initializer="normal", relation_initializer="normal")
loss		BCEWithLogitsLoss
training_kwargs	dict(num_epochs=25, checkpoint_-name="checkpoint_name.pt", checkpoint_frequency=0)	dict(num_epochs=25, checkpoint_-name="checkpoint_name.pt", checkpoint_frequency=0)
random_seed	5678	5678

Table 4.4 **Keyword arguments provided to PyKEEN's pipeline object during model training.**

score of the most likely tail entities to complete those relations ("predict target"). To compare with the RF implementation, we used the same test set triples with option (2), or the predict triples method. The PyKEEN implementation requires a relation type to be specified for all triples with this method, so for the negative triples, we randomly sampled the relation type out of the available types (desiccation, drought, or both). The RESCAL authors state that "link prediction can be done by comparing [the plausibility score] to some given threshold" [Nickel et al., 2011]. Practically, RESCAL is a tensor factorization graph embedding model, where a loss function is used to optimize the model during training. The choice of loss function determines how the triple scores produced by RESCAL can be interpreted. The default loss function for PyKEEN's implementation is a pairwise loss function, which takes one negative and one positive triple at a time, and optimizes such that the positive triple should always receive a higher score than the negative triple. The result of this optimization method is that there is no global threshold around which positive and negative triple scores are optimized. The score of a positive triple from one pair could be less than the score of a negative triple from another pair. Therefore, classification of triples by defining a score threshold as suggested in [Nickel et al., 2011] is not possible when the algorithm is optimized with a pairwise loss. On the other hand, pointwise losses do optimize around a global threshold of 0; triples with scores greater than 0 should be positive, while those with scores less than 0 should be negative. In our case, the default loss function for RESCAL in PyKEEN (MarginRankingLoss) is a pairwise loss function, which means that we could not use the RESCAL models trained with

MarginRankingLoss to perform triple classification. Instead, we needed to use a pointwise loss function, of which BCEWithLogitsLoss is one. Using the same RESCAL model trained with BCEWithLogitsLoss as we did for the RF models above, we asked RESCAL to provide plausibility scores for the 4,000 test set triples. Using 0 as a classification threshold, we calculated an F1 score for this method as a classifier.

In addition to evaluating the ability of the model to correctly classify a set of triples, we also used the "predict all" functionality to assess the model's capability to identify high-probability triples, keeping the top 100 triples. We manually assessed the validity of the top 10 relations using a Web of Science search. For each pair, we performed a search for the query <head entity> AND <tail entity>. For entities with specific terms, we simplified the entity and performed an additional search; an example is "wheat germ systems", which was changed to "wheat germ".

Temporal link prediction

To determine a suitable algorithm for TLP on our network, we first performed a literature search on available model architectures for TLP on heterogeneous networks (**Table S4.2**). However, while we were able to identify several unique algorithms, only six of these had code associated with the paper; of these six, only only two have code that has been updated less than four years ago, and of those two, only one had code that was serviceably documented enough to use without major modification. We chose the algorithm with recent and serviceably reusable code, STHN, for use in our experiments. After communicating with the developer to clarify details regarding input data formatting, we ran the STHN algorithm with –max_edges set to 50, and with the –predict_class option.

Code availability

All code for this project is available at https://github.com/serenalotreck/literature-genes.

CHAPTER 5

CONCLUSION

Link prediction on literature-derived KG for hypothesis generation leaves room to grow

In this work, I demonstrated that using literature-derived KG with static link prediction methods is insufficient to provide high-quality automatic hypothesis generation. However, any study involving the implementation of specific methods to solve a problem is limited by the time and imagination of the study designer, and there are always more options and methods to try. Therefore, although this thesis does not present promising results regarding hypothesis generation on KG, results are limited to a specific subset of methods, and there is much room for future improvement.

Having intended to work on both KG construction and link prediction in this thesis, it is my opinion that the two tasks each deserve their own separate investigations. KG construction and evaluation is a labor-intensive process, and shown in this work, link prediction doesn't offer a guarantee of useful predictions. For a potential future PhD student, I would therefore suggest focusing on either KG construction, or link prediction/hypothesis generation on an existing, high-quality KG, as an appropriate course of action likely to maximize success.

For those interested in pursuing the KG construction avenue of future work, there are three areas that I believe would benefit from special focus. The first is the problem setup for KG construction. The motivation for using literature as a data source of KG is that existing database-derived graphs are limited in scope due to the manual curation requirements of the databases from which they are built. However, using literature alone to build a KG also result in an incomplete graph due to the limitations of information extraction, as well as the difficulty of resolving entity mentions with varied spellings or synonyms to create a graph that is robust and easily usable. Therefore, I would suggest focusing construction efforts on using the literature to complete an existing database-derived graph. There are many stellar examples of KG in both the plant sciences and beyond, based predominantly on database sources, as discussed in the Introduction. Many of the larger, better-established KG benefit from entity grounding, where each node and edge is linked to a unique identifier, and mentions of the same underlying, real-world object can be resolved. Completing

a database graph will require careful thought about the study system in question as it relates to the content of the graph, as most existing KG are built around model or crop species. It will also require careful thought about how to successfully integrate literature-derived data into the KG. Entity resolution was completely excluded from this thesis due to logistical constraints; however, I believe that this work would have benefited greatly from investing time and resources into building entity resolution into our pipeline. By focusing on using literature to complete an existing high-quality KG, the final product will likely be more useful for future efforts in hypothesis generation than either a database-only or literature-only KG.

The second area for focus in a KG construction project should be the information extraction methods used to draw entities and relations from the literature. As seen in this thesis, current methods struggle especially to identify relations in text, which are a fundamental component of creating a useful KG. Creation of a gold standard dataset, like that described in the first chapter of this thesis, specifically for the domain in which the KG is being created, will be very important for both evaluation of entity/relation extraction methods. Many improvements can be made on the annotation procedures presented in the first chapter of this thesis (the PICKLE dataset) given greater time and labor inputs. My two principal recommendations would be to (a) use defined ontologies for entity annotations and to (b) include entity resolution in annotations. The use of defined ontologies, as demonstrated in the CRAFT corpus' annotation guidelines [Bada et al., 2012], goes a long way to improving inter-annotator agreement, and has the added benefit of improving the ease with which entities can be integrated into a KG drawn from databases using the same ontologies. Entity resolution, as mentioned above, will likely improve the quality of the resulting KG, as well as potentially improving the performance of entity and relation extraction algorithms, as shown in [Wadden et al., 2019]. In addition to these changes in annotation guidelines, I would also recommend annotating as large a volume of documents as possible. While we found that the performance of the DyGIE++ algorithm did not improve on relation and entity extraction after $\tilde{1}50$ documents in the training set for the PICKLE dataset, a larger volume of documents in the annotation set will likely make evaluations of performance and KG quality more robust. In

addition to designing a larger and higher-quality training and evaluation annotated dataset, it will be important to consider approaches that can bridge sentence boundaries to extract relations. One likely weakness of the DyGIE++ for our use-case is that, like many other relation extraction methods, it is limited to extracting relations at the sentence level for reasons of computational complexity. Important biological relations may not be directly stated in a single sentence; therefore, methods that can perform document-level relationship extraction will likely help built better KG. That being said, poor relation extraction performance extended to the document-level GPT-based method that we employed. Luckily, there is always up to go with prompt engineering, and this is a ripe area for future research.

If the future researcher is able to successfully improve relation extraction to the point where almost all relevant relations are being recovered from scientific abstracts, my third recommendation would be to consider the literature data source being used to build the KG. One of the most common questions I've received about my work on KG is why we are using abstracts instead of full text articles. There are two reasons: first, the ease of access, as there is no guarantee of access to full text in a workable (non-PDF) format for full text articles. Abstracts are much more accessible than full text; as of June 20204, there were 37 million citations in PubMed, but as of fiscal year 2023, PubMed Central only contained 9,407,149 articles [PubMed, 2024]. While many institutions have access to paywalled articles, high-throughput collection and processing of these articles is non-trivial. The second justification for using abstracts is that in principle, the most important findings of a paper should be described succinctly in an abstract. However, based on our findings in the previous chapter, the abstract alone does not appear to be sufficiently information-rich to build a high quality KG. It is possible that full text can provide better information for KG construction: one paper quantified the difference in biomedical entities in PubMed abstracts versus free full text available on PubMed Central [Müller et al., 2010]. The team found that on average, about 10% of entities are only found in abstracts, while 75 - 86% of entities are only found in the full text. Another past study showed that while information density was high in abstracts (more unique entities per length of text), the information coverage of the full text was much greater [Schuemie et al., 2004].

While it is clear that the full text is more information-rich for entities, I was unable to find any similar work on the richness of relationships; this would be an excellent area for future study. However, since there are so many more entities in the full text, we can likely assume that there are more relations in full text as well. Therefore, if the performance of relation extraction methods is relatively assured, finding ways to incorporate full text into KG construction for biological domains will likely improve the quality of the resulting KG.

For a researcher more interested in extending the hypothesis generation aspect of this work, my principal recommendation is to evaluate the landscape of project motivation via a more human-centered approach, evaluating the needs of researchers in order to inform a more appropriate hypothesis generation solution. Rather than assuming that the desired result of a hypothesis generation system is a fully automated process that removes the human researcher from the loop, as was this case in this work, further studies involving human participants could provide insight into more nuanced and potentially more effective forms of hypothesis generation. Domain expertise is still necessary in a world with automated hypothesis generation, and we often gain our domain expertise by reading the literature during the process of manual hypothesis generation. In addition, we already posses a great number of tools designed to augment researchers' efficiency in searching the literature – are there other modes of augmentation, in other parts of the manual hypothesis generation pipeline, that researchers might want? Are KG the right tool to accomplish the observed goals of real people, or is there some other entirely different avenue down which this work could or should progress? In the Introduction of this thesis, I discuss tools such as AgroLD and KnetMiner, which are large KG designed to help scientists explore the known landscape of plant biology. As essentially an information scientist, I view such tools as well-developed and very useful to biologists. However, asking around in my own community, I have yet to encounter a biologist who is aware of such KG-based tools before I describe them. An area of future research that I think is very important is the use of surveys to engage with potential stakeholders to investigate potential synergies between how they currently generate hypotheses and KG-based or other tools for automated hypothesis generation. Given the connection between domain expertise and hypothesis

generation, it is unlikely that a fully automated system would make sense for scientists; therefore, further work is required to determine what the needs are of the scientific community in regards to hypothesis generation. Additionally, involvement of domain experts once the hypothesis generation method has been determined is extremely important. The test set for a link prediction model, for example, while useful for evaluating model performance in a vacuum, cannot tell you whether or not the links you are predicting are meaningful. If a domain expert doesn't find the connections between organisms and chemical elements being predicted by the model relevant, it doesn't matter that the system is good at predicting them.

In terms of the technical aspect of hypothesis generation, I would express caution in regards to continuing using KG embedding methods for link prediction as the primary mode of hypothesis generation. My intuition says that, while improvement in the quality of the underlying KG and incorporation of a temporal component will likely yield some improvement in link prediction capabilities, the low baseline performance of the static link prediction methods on a middle-quality graph seen here cause me to suspect that any performance increases may still not be sufficient to generate actionable hypotheses. My experience of working with KG embedding models was spending a lot of time trying different hyperparameters, inputs, and models, for very little corresponding improvement in performance. While I have certainly not exhaustively tried every option that even just the PyKEEN package implements, and the graphs on which I was working were not of the highest-quality, I can envision a scenario where such an exhaustive search doesn't provide any meaningful improvement over the initial implementation. As discussed briefly in the conclusion of the previous chapter, investigation of the impacts of network structure on embedding algorithms could potentially help illuminate why link prediction models fail on biological datasets. I would recommend a thorough evaluation of the impact of degree skew (scale-free or approximation of scale-free behavior) on embedding models, especially as compared to the datasets like the Kinships dataset on which models like RESCAL are often evaluated. Additionally, on a higher level, thorough exploration of the interests of the research community for automated or augmented hypothesis generation methods will provide insight into potential alternative methods.

KG-based or otherwise, there likely exist methods that could have a much greater degree of success than link prediction. It seems of the utmost importance to me to explore possible other avenues for hypothesis generation before resuming work on the link prediction trajectory presented in this thesis.

A brief note on the role of information overload in the writing of this dissertation and its implications

As noted by [Bawden and Robinson, 2020], even when writing on information overload, information overload remains a problem, and selectivity in citations is necessary to maintain focus. When I first began this research in 2019, my launching point into the field of hypothesis generation was through knowledge graph completion. As a result of only using search terms related to knowledge graphs while reading the literature to propose my dissertation research, I developed a kind of literature myopia, where the selectivity of my citations was biased away from the broader field in which my research was situated; this became even more problematic upon what is referred to in systematic literature reviews as “backwards search", where the researcher follows the citations in their initial search results [Foo et al., 2021, Xiao and Watson, 2019]. The extent of this myopia only became clear when the methods I had identified as promising candidates for hypothesis generation started failing; I returned to the literature with the sense that I had maybe missed something important. I decided to step back with the specificity of my search terms and instead of knowledge graphs, simply search the phrase "automated hypothesis generation". Since then, I have luckily stumbled upon several other terms that seem to encompass the body of literature in this field, and would like to explicitly state them here: automated hypothesis generation, automatic hypothesis generation, and literature-based discovery. While using a diversity of search terms may seem obvious, the knowledge of exactly which terms to search to gain a comprehensive understanding of the state of this field, which I will predominantly refer to as automated hypothesis generation, took me several years to come to. In addition, even the seemingly trivial difference between "automated" and "automatic" in a search engine drastically changed the papers that turned up in the results, which is why I feel it is important to point out just how dramatically a bias in search terms can

affect the process of science for a given individual.

Existing work has demonstrated the effect of word choice in titles and abstracts on the visibility of papers in search engine results, indicating that papers that use more jargon are less often cited [Martínez and Mammola, 2020], and there exists a body of literature containing recommendations for search engine optimization through considered formulation of titles and abstracts to change the overall visibility of an academic paper [Shahzad et al., 2017, Pottier et al., 2023]. There is an additional body of literature containing recommendations on how to formulate systemic reviews, include recommendations on how to choose search terms for systematic reviews. However, these papers predominantly assume that researchers are aware of the possible search terms they would need, and the advice is catered towards choosing search terms out of a known set, including advice like "The keywords for the search should be derived from the research question(s)" [Xiao and Watson, 2019], "A major consideration in systematic searching is balancing the principles of sensitivity and specificity" [Purssell and McCrae, 2020], or even just using an example search term without explaining how it was chosen [Foo et al., 2021]. These recommendations contain useful advice such as expanding your search terms by abbreviations, and in the case of [Xiao and Watson, 2019], even address the issue of alternative terms:

> "Second, researchers doing cross-country studies should pay attention to the cultural difference in terminology. For instance, “eminent domain” is called “compulsory acquisition” and “parking lot” called “car park” in Australia and New Zealand. “Urban revitalization” is typically called “urban regeneration” in the United Kingdom. The search can only be successful if we use the correct vocabulary from the culture of study. Third, Bayliss and Beyer (2015) brought up the issue of the evolving vocabulary. For example, the interstate highway system was originally called “interstate and defense highways” because it was constructed for defense purposes in the cold war era (Weingroff 1996). The term “defense” was then dropped from the name. Therefore, researchers should be conscious of the vocabulary changes over time. In the search of literature dated back in history, one should use the correct vocabulary from that period

of time."

However, note that the phrasing of this advice implies a pre-existing knowledge of alternative terms in a field. In scientific fields that do not necessitate a geographic focus in the same way urban planning does, nearly all literature searches are international by default. Disciplines can be divided terminologically along invisible boundaries that don't correspond to something evident like geography or time, and to the best of my knowledge, there is no body of work that has quantified the effect of this invisible prerequisite knowledge on systematic literature reviews or citation metrics. As a result of my experiences in writing this dissertation, my personal definition of information overload has now expanded to include the process by which important information is effectively hidden from an individual because they do not already possess some invisible prerequisite knowledge. My personal experience was that performing a research project outside my lab's expertise meant that I didn't have an inside source who was aware of the various terminology that I needed to search; however, given the lack of quantification of this phenomenon, who is to say that experienced researchers are unknowingly missing important or novel literature in their field as a result of terminology differences? Because search terms determine so much of how we process and share information as scientists, I would be extremely interested to see the results of future work exploring the impact of invisible prerequisite knowledge on bibliometrics like those explored in the third chapter of this thesis. Additionally, because the hypothesis generation system explored in this thesis as well as many other potential approaches rely on the output of a scientific literature search, quantification of the impact of missing search terms will be important to the efficacy of potential future literature-based hypothesis generation systems.

BIBLIOGRAPHY

[Abu-Salih et al., 2023] Abu-Salih, B., AL-Qurishi, M., Alweshah, M., AL-Smadi, M., Alfayez, R., and Saadeh, H. (2023). Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities. *Journal of Big Data*, 10(1):81.

[Akujuobi, 2021] Akujuobi, U. (2021). Revolutionizing Hypothesis Generation.

[Alger, 2019] Alger, B. E. (2019). The Scientific Hypothesis Today. In Alger, B. E., editor, *Defense of the Scientific Hypothesis: From Reproducibility Crisis to Big Data*, page 0. Oxford University Press.

[Ali et al., 2021] Ali, M., Berrendorf, M., Hoyt, C. T., Vermue, L., Sharifzadeh, S., Tresp, V., and Lehmann, J. (2021). PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings.

[Alpert, 2005] Alpert, P. (2005). The Limits and Frontiers of Desiccation-Tolerant Life. *Integrative and Comparative Biology*, 45(5):685–695.

[Angeli et al., 2015] Angeli, G., Johnson Premkumar, M. J., and Manning, C. D. (2015). Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

[Bada et al., 2012] Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., and Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):161.

[Bawden and Robinson, 2020] Bawden, D. and Robinson, L. (2020). Information Overload: An Introduction. In *Oxford Research Encyclopedia of Politics*. Oxford University Press.

[Bekhuis, 2006] Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, 3(1):2.

[Bewley, 1979] Bewley, J. D. (1979). Physiological Aspects of Desiccation Tolerance. *Annual Review of Plant Physiology*, 30(1):195–238. _eprint: https://doi.org/10.1146/annurev.pp.30.060179.001211.

[Bian et al., 2019] Bian, R., Koh, Y. S., Dobbie, G., and Divoli, A. (2019). Network Embedding and Change Modeling in Dynamic Heterogeneous Networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 861–864, Paris France. ACM.

[Bleker et al., 2023] Bleker, C., Ramšak, Z., Bittner, A., Podpečan, V., Zagorščak, M., Wurzinger, B., Baebler, , Petek, M., Križnik, M., Dieren, A. v., Gruber, J., Afjehi-Sadat, L., Županič, A., Teige, M., Vothknecht, U. C., and Gruden, K. (2023). Stress Knowledge Map: A knowledge graph resource for systems biology analysis of plant stress responses. Pages: 2023.11.28.568332 Section: New Results.

[Cachola et al., 2020] Cachola, I., Lo, K., Cohan, A., and Weld, D. (2020). TLDR: Extreme Summarization of Scientific Documents. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777. Conference Name: Findings of the Association for Computational Linguistics: EMNLP 2020 Place: Online Publisher: Association for Computational Linguistics.

[Cai et al., 2023] Cai, B., Xiang, Y., Gao, L., Zhang, H., Li, Y., and Li, J. (2023). Temporal Knowledge Graph Completion: A Survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6545–6553. arXiv:2201.08236 [cs].

[Cai et al., 2018] Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

[Caufield et al., 2024] Caufield, J. H., Hegde, H., Emonet, V., Harris, N. L., Joachimiak, M. P., Matentzoglu, N., Kim, H., Moxon, S., Reese, J. T., Haendel, M. A., Robinson, P. N., and Mungall, C. J. (2024). Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3):btae104.

[Chen et al., 2019] Chen, H., Cao, G., Chen, J., and Ding, J. (2019). A Practical Framework for Evaluating the Quality of Knowledge Graph. In Zhu, X., Qin, B., Zhu, X., Liu, M., and Qian, L., editors, *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding*, Communications in Computer and Information Science, pages 111–122, Singapore. Springer.

[Cooper et al., 2024] Cooper, L., Elser, J., Laporte, M.-A., Arnaud, E., and Jaiswal, P. (2024). Planteome 2024 Update: Reference Ontologies and Knowledgebase for Plant Biology. *Nucleic Acids Research*, 52(D1):D1548–D1555.

[Darnala et al., 2023] Darnala, B., Amardeilh, F., Roussey, C., Todorov, K., and Jonquet, C. (2023). C3PO: a crop planning and production process ontology and knowledge graph. *Frontiers in Artificial Intelligence*, 6. Publisher: Frontiers.

[Dileo et al., 2023] Dileo, M., Zignani, M., and Gaito, S. (2023). DURENDAL: Graph deep learning framework for temporal heterogeneous networks. arXiv:2310.00336 [cs].

[Fo et al., 2023] Fo, K., Chuah, Y. S., Fyh, H., Davey, E. E., Fullwood, M., Thibault, G., and Mutwil, M. (2023). PlantConnectome: knowledge networks encompassing >100,000 plant article abstracts. Pages: 2023.07.11.548541 Section: New Results.

[Foo et al., 2021] Foo, Y. Z., O'Dea, R. E., Koricheva, J., Nakagawa, S., and Lagisz, M. (2021). A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution. *Methods in Ecology and Evolution*, 12(9):1705–1720. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13654.

[Grzyb and Skłodowska, 2022] Grzyb, T. and Skłodowska, A. (2022). Introduction to Bacterial Anhydrobiosis: A General Perspective and the Mechanisms of Desiccation-Associated Damage. *Microorganisms*, 10(2):432. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[Hassani-Pak et al., 2016] Hassani-Pak, K., Castellote, M., Esch, M., Hindle, M., Lysenko, A., Taubert, J., and Rawlings, C. (2016). Developing integrated crop knowledge networks to advance candidate gene discovery. *Applied & Translational Genomics*, 11:18–26.

[Hassani-Pak et al., 2021] Hassani-Pak, K., Singh, A., Brandizi, M., Hearnshaw, J., Parsons, J. D., Amberkar, S., Phillips, A. L., Doonan, J. H., and Rawlings, C. (2021). KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnology Journal*, 19(8):1670–1678. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbi.13583.

[Hibshman et al., 2020] Hibshman, J. D., Clegg, J. S., and Goldstein, B. (2020). Mechanisms of Desiccation Tolerance: Themes and Variations in Brine Shrimp, Roundworms, and Tardigrades. *Frontiers in Physiology*, 11.

[Imbert et al., 2023] Imbert, B., Kreplak, J., Flores, R.-G., Aubert, G., Burstin, J., and Tayeh, N. (2023). Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case on grain legumes. *Frontiers in Artificial Intelligence*, 6. Publisher: Frontiers.

[Islam et al., 2021] Islam, M. K., Aridhi, S., and Smaïl-Tabbone, M. (2021). Simple negative sampling for link prediction in knowledge graphs. In *The 10th International Conference on Complex Networks and their Applications*, volume 1016, pages 549–562, Madrid, Spain. Springer International Publishing.

[Issa et al., 2021] Issa, S., Adekunle, O., Hamdi, F., Cherfi, S. S.-S., Dumontier, M., and Zaveri, A. (2021). Knowledge Graph Completeness: A Systematic Literature Review. *IEEE Access*, 9:31322–31339. Conference Name: IEEE Access.

[Kanehisa, 2002] Kanehisa, M. (2002). The KEGG Database. In *'In Silico' Simulation of Biological Processes*, pages 91–103. John Wiley & Sons, Ltd. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470857897.ch8.

[Klerings et al., 2015] Klerings, I., Weinhandl, A. S., and Thaler, K. J. (2015). Information overload in healthcare: too much of a good thing? *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 109(4):285–290.

[Kong et al., 2019] Kong, C., Li, H., Zhang, L., Zhu, H., and Liu, T. (2019). Link Prediction on Dynamic Heterogeneous Information Networks. In Tagarelli, A. and Tong, H., editors, *Computational Data and Social Networks*, pages 339–350, Cham. Springer International Publishing.

[Landhuis, 2016] Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, 535(7612):457–458. Number: 7612 Publisher: Nature Publishing Group.

[Larmande and Todorov, 2021] Larmande, P. and Todorov, K. (2021). AgroLD: A Knowledge Graph for the Plant Sciences. In Hotho, A., Blomqvist, E., Dietze, S., Fokoue, A., Ding, Y., Barnaghi, P., Haller, A., Dragoni, M., and Alani, H., editors, *The Semantic Web – ISWC 2021*, volume 12922, pages 496–510. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

[Le Guillarme and Thuiller, 2022] Le Guillarme, N. and Thuiller, W. (2022). TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3):625–641. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13778.

[León-Lobos et al., 2012] León-Lobos, P., Way, M., Aranda, P. D., and Lima-Junior, M. (2012). The role of ex situ seed banks in the conservation of plant diversity and in ecological restoration in Latin America. *Plant Ecology & Diversity*, 5(2):245–258.

[Li et al., 2023] Li, C., Hong, R., Xu, X., Trajcevski, G., and Zhou, F. (2023). Simplifying Temporal Heterogeneous Network for Continuous-Time Link prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1288–1297, Birmingham United Kingdom. ACM.

[Lotreck et al., 2023] Lotreck, S., Segura Abá, K., Lehti-Shiu, M. D., Seeger, A., Brown, B. N. I., Ranaweera, T., Schumacher, A., Ghassemi, M., and Shiu, S.-H. (2023). Plant Science Knowledge Graph Corpus: a gold standard entity and relation corpus for the molecular plant sciences. *in silico Plants*, 6(1):diad021.

[Lotreck et al., 2024] Lotreck, S. G., Ghassemi, M., and VanBuren, R. T. (2024). Unifying the research landscape of desiccation tolerance to identify trends, gaps, and opportunities. *bioRxiv*.

[Marks et al., 2021] Marks, R. A., Farrant, J. M., Nicholas McLetchie, D., and VanBuren, R. (2021). Unexplored dimensions of variability in vegetative desiccation tolerance. *American Journal of Botany*, 108(2):346–358. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajb2.1588.

[Martínez and Mammola, 2020] Martínez, A. and Mammola, S. (2020). Specialized terminology limits the reach of new scientific knowledge. Pages: 2020.08.20.258996 Section: New Results.

[Milošević and Thielemann, 2023] Milošević, N. and Thielemann, W. (2023). Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *Journal of Web Semantics*, 75:100756.

[Müller et al., 2010] Müller, B., Klinger, R., Gurulingappa, H., Mevissen, H.-T., Hofmann-Apitius, M., Fluck, J., and Friedrich, C. M. (2010). Abstracts versus Full Texts and Patents: A Quantitative Analysis of Biomedical Entities. In Cunningham, H., Hanbury, A., and Rüger, S., editors, *Advances in Multidisciplinary Retrieval*, pages 152–165, Berlin, Heidelberg. Springer.

[Ni et al., 2023] Ni, X., Zhao, Y., and Yao, Y. (2023). Dynamic Heterogeneous Link Prediction Based on Hierarchical Attention Model. In *Proceedings of the 8th International Conference on Cyber Security and Information Engineering*, pages 111–115, Putrajaya Malaysia. ACM.

[Nicholson and Greene, 2020] Nicholson, D. N. and Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18:1414–1428. Publisher: Elsevier.

[Nickel et al., 2011] Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A Three-Way Model for Collective Learning on Multi-Relational Data.

[Peng et al., 2023] Peng, C., Xia, F., Naseriparsa, M., and Osborne, F. (2023). Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, 56(11):13071–13102.

[Persson et al., 2011] Persson, D., Halberg, K. A., Jørgensen, A., Ricci, C., Møbjerg, N., and Kristensen, R. M. (2011). Extreme stress tolerance in tardigrades: surviving space conditions in low earth orbit. *Journal of Zoological Systematics and Evolutionary Research*, 49(s1):90–97. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1439-0469.2010.00605.x.

[Pottier et al., 2023] Pottier, P., Lagisz, M., Burke, S., Drobniak, S. M., Downing, P. A., Macartney, E. L., Martinig, A. R., Mizuno, A., Morrison, K., Pollo, P., Ricolfi, L., Tam, J., Williams, C., Yang, Y., and Nakagawa, S. (2023). Keywords to success: a practical guide to maximise the visibility and impact of academic papers.

[PubMed, 2024] PubMed (2024). About PMC.

[Purssell and McCrae, 2020] Purssell, E. and McCrae, N. (2020). Searching the Literature. In Purssell, E. and McCrae, N., editors, *How to Perform a Systematic Literature Review: A Guide for Healthcare Researchers, Practitioners and Students*, pages 31–44. Springer International Publishing, Cham.

[Qin and Yeung, 2024] Qin, M. and Yeung, D.-Y. (2024). Temporal Link Prediction: A Unified Framework, Taxonomy, and Review. *ACM Computing Surveys*, 56(4):1–40.

[Ramšak et al., 2018] Ramšak, , Coll, A., Stare, T., Tzfadia, O., Baebler, S., Van De Peer, Y., and Gruden, K. (2018). Network Modeling Unravels Mechanisms of Crosstalk between Ethylene and Salicylate Signaling in Potato. *Plant Physiology*, 178(1):488–499.

[Raymond, 2019] Raymond, D. (2019). Using Artificial Intelligence to Combat Information Overload in Research. *IEEE Pulse*, 10(1):18–21. Conference Name: IEEE Pulse.

[Rossi et al., 2021] Rossi, A., Firmani, D., Matinata, A., Merialdo, P., and Barbosa, D. (2021). Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. *ACM Transactions on Knowledge Discovery from Data*, 15(2):1–49. arXiv:2002.00819 [cs, stat].

[Sajadmanesh et al., 2019] Sajadmanesh, S., Bazargani, S., Zhang, J., and Rabiee, H. R. (2019). Continuous-Time Relationship Prediction in Dynamic Heterogeneous Information Networks. *ACM Transactions on Knowledge Discovery from Data*, 13(4):1–31.

[Schuemie et al., 2004] Schuemie, M. J., Weeber, M., Schijvenaars, B. J. A., Van Mulligen, E. M., Van Der Eijk, C. C., Jelier, R., Mons, B., and Kors, J. A. (2004). Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604.

[Seo et al., 2022] Seo, S., Cheon, H., Kim, H., and Hyun, D. (2022). Structural Quality Metrics to Evaluate Knowledge Graphs. arXiv:2211.10011 [cs].

[Sett et al., 2018] Sett, N., Basu, S., Nandi, S., and Singh, S. R. (2018). Temporal link prediction in multi-relational network. *World Wide Web*, 21(2):395–419.

[Shahzad et al., 2017] Shahzad, A., Mohd Nawi, N., Abd Hamid, N., Khan, S. N., Aamir, M., Ullah, A., and Abdullah, S. (2017). The Impact of Search Engine Optimization on The Visibility of Research Paper and Citations. *JOIV : International Journal on Informatics Visualization*, 1(4-2):195–198.

[Smalheiser, 2012] Smalheiser, N. R. (2012). Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science and Technology*, 63(2):218–224.

[Smalheiser and Swanson, 1998] Smalheiser, N. R. and Swanson, D. R. (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3):149–153.

[Swanson, 1986] Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18. Publisher: Johns Hopkins University Press.

[The Gene Ontology Consortium, 2019] The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338.

[Thessen et al., 2023] Thessen, A. E., Cooper, L., Swetnam, T. L., Hegde, H., Reese, J., Elser, J., and Jaiswal, P. (2023). Using knowledge graphs to infer gene expression in plants. *Frontiers in Artificial Intelligence*, 6. Publisher: Frontiers.

[Unni et al., 2022] Unni, D. R., Moxon, S. A. T., Bada, M., Brush, M., Bruskiewich, R., Caufield, J. H., Clemons, P. A., Dancik, V., Dumontier, M., Fecho, K., Glusman, G., Hadlock, J. J., Harris, N. L., Joshi, A., Putman, T., Qin, G., Ramsey, S. A., Shefchek, K. A., Solbrig, H., Soman, K., Thessen, A. E., Haendel, M. A., Bizon, C., Mungall, C. J., and The Biomedical Data Translator Consortium (2022). Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and Translational Science*, 15(8):1848–1855.

[Wadden et al., 2019] Wadden, D., Wennberg, U., Luan, Y., and Hajishirzi, H. (2019). Entity, Relation, and Event Extraction with Contextualized Span Representations. arXiv:1909.03546 [cs].

[Wang et al., 2021a] Wang, M., Ma, X., Si, J., Tang, H., Wang, H., Li, T., Ouyang, W., Gong, L., Tang, Y., He, X., Huang, W., and Liu, X. (2021a). Adverse Drug Reaction Discovery Using a Tumor-Biomarker Knowledge Graph. *Frontiers in Genetics*, 11. Publisher: Frontiers.

[Wang et al., 2021b] Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., Wu, T., and Chen, H. (2021b). Knowledge graph quality control: A survey. *Fundamental Research*, 1(5):607–626.

[Wren, 2008] Wren, J. (2008). The 'Open Discovery' Challenge. pages 39–55.

[Xiao and Watson, 2019] Xiao, Y. and Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research*, 39(1):93–112. Publisher: SAGE Publications Inc.

[Xue et al., 2020] Xue, H., Yang, L., Jiang, W., Wei, Y., Hu, Y., and Lin, Y. (2020). Modeling Dynamic Heterogeneous Network for Link Prediction using Hierarchical Attention with Temporal RNN. arXiv:2004.01024 [cs, stat].

[Yin et al., 2019] Yin, Y., Ji, L.-X., Zhang, J.-P., and Pei, Y.-L. (2019). DHNE: Network Representation Learning Method for Dynamic Heterogeneous Networks. *IEEE Access*, 7:134782–134792.

[Yue et al., 2022] Yue, C., Du, L., Fu, Q., Bi, W., Liu, H., Gu, Y., and Yao, D. (2022). HTGN-BTW: Heterogeneous Temporal Graph Network with Bi-Time-Window Training Strategy for Temporal Link Prediction. *ArXiv*.

[Zhou et al., 2018] Zhou, L., Yang, Y., Ren, X., Wu, F., and Zhuang, Y. (2018). Dynamic Network Embedding by Modeling Triadic Closure Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). Number: 1.

APPENDIX

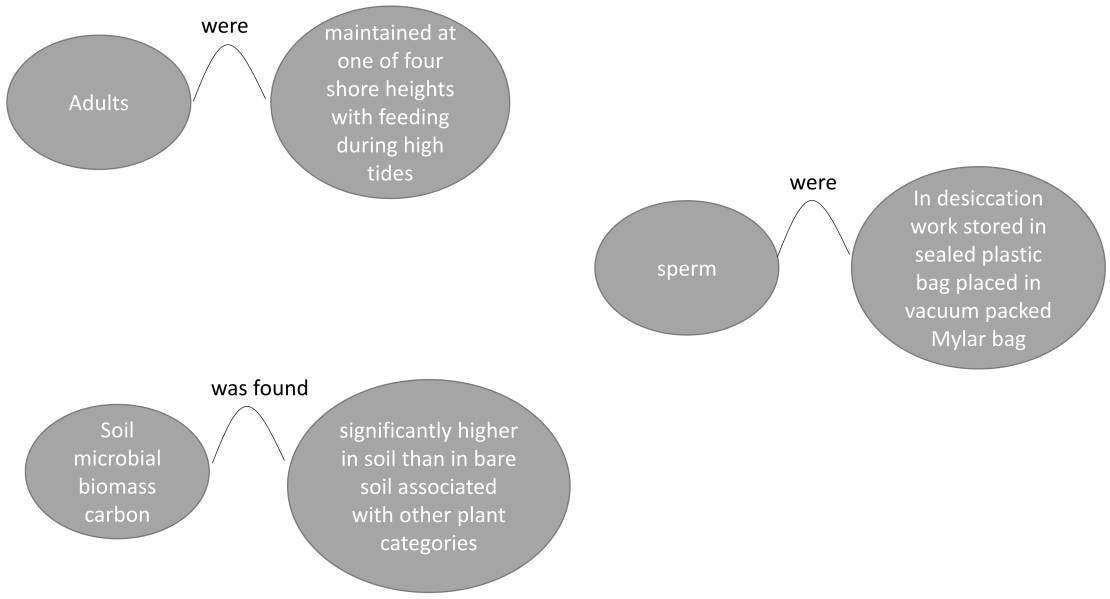


Figure S4.1 **Examples of low-quality OpenIE triples.** Real triples extracted from the dataset using OpenIE. Because OpenIE is schema-free and domain-agnostic, it can only rely on syntactic (grammatical) rules, and therefore extracts extremely long clauses as entities, leading to uninformative relations.

Head entity	Edge type	Tail entity	AND search
wheat germ system	both	deciduous seasonal forests	(TS=(wheat germ)) AND TS=(deciduous forests)
opuntia fragilis	both	deciduous seasonal forests	(TS=(opuntia fragilis)) AND TS=(deciduous forests)
pshsfa7a1_2595	both	mwsp	(TS=(pshsfa7a1)) AND TS=(mwsp)
salt stress-induced calcium signal	both	c. stelligera	(TS=(calcium)) AND TS=(c. stelligera)
bzip23 transcription factor activity	both	deciduous seasonal forests	(TS=(bzip)) AND TS=(deciduous forests)
beech-fir stand	both	deciduous seasonal forests	(TS=(beech-fir stand)) AND TS=(deciduous forests)
t. fluminensis	both	deciduous seasonal forests	(TS=(Tradescantia fluminensis)) AND TS=(deciduous forests)
crocus sativus l	both	reds/breaker tomatoes	(TS=(crocus sativus l)) AND TS=(tomatoes)
drought-responsive and jasmonic acid biosynthesis genes	both	b. amphitrite	(TS=(jasmonic acid)) AND TS=(Amphibalanus amphitrite)
mandarin water	both	deciduous seasonal forests	(TS=(mandarin water)) AND TS=(deciduous forests)

Table S4.1 **Final search queries for top ten predicted triples.**

Method	Reference	Year	Code available?
DynamicTriad	[Zhou et al., 2018]	2018	Yes
TMLP	[Sett et al., 2018]	2018	No
NP-GLM	[Sajadmanesh et al., 2019]	2019	No
DHNE	[Yin et al., 2019]	2019	Yes
HA-LSTM	[Kong et al., 2019]	2019	No
Change2vec	[Bian et al., 2019]	2019	Yes
DyHATR	[Xue et al., 2020]	2020	Yes
HTGN-BTW	[Yue et al., 2022]	2022	No
Att-ConvLSTM	[Ni et al., 2023]	2023	No
STHN	[Li et al., 2023]	2023	Yes
DURENDAL	[Dileo et al., 2023]	2023	Yes

Table S4.2 **Summary of literature search for heterogeneous TLP methods.**

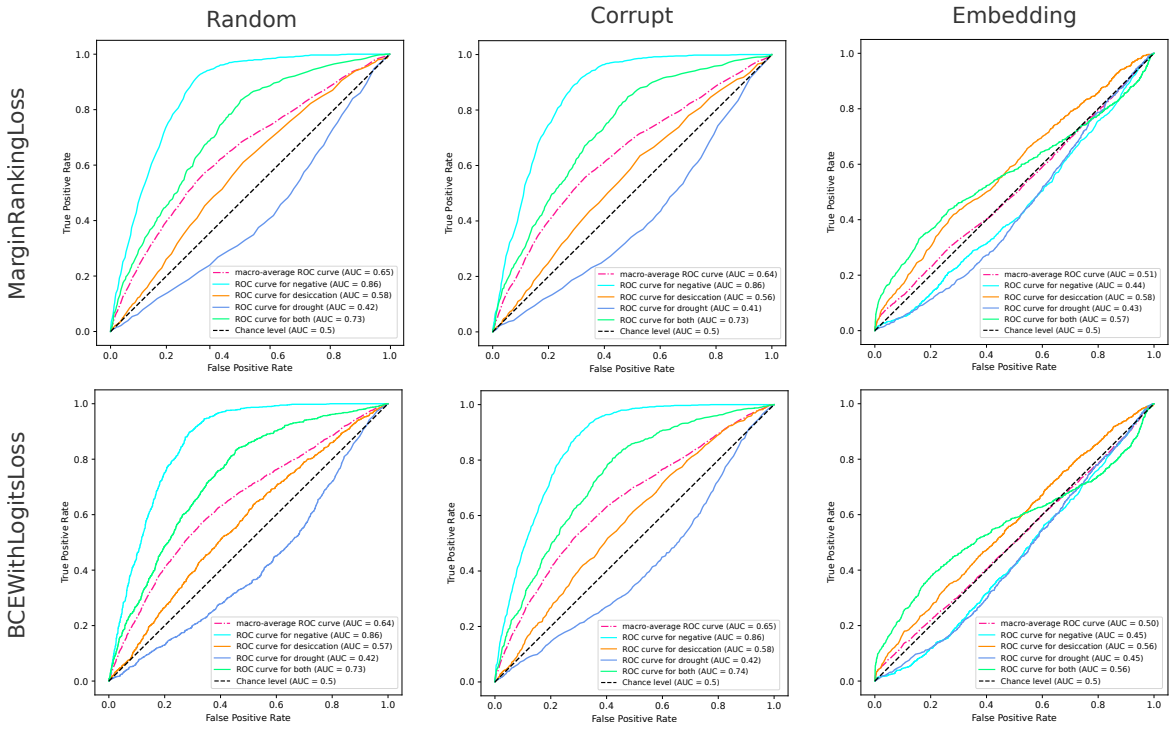


Figure S4.2 **AUROC scores for RF models with different sampling strategies.** AUROC curves for models for each upstream RESCAL loss function and negative sampling strategy. No meaningful difference was found between the two RESCAL losses in the RF models, and there was no meaningful difference between the random and corrupt sampling strategies.

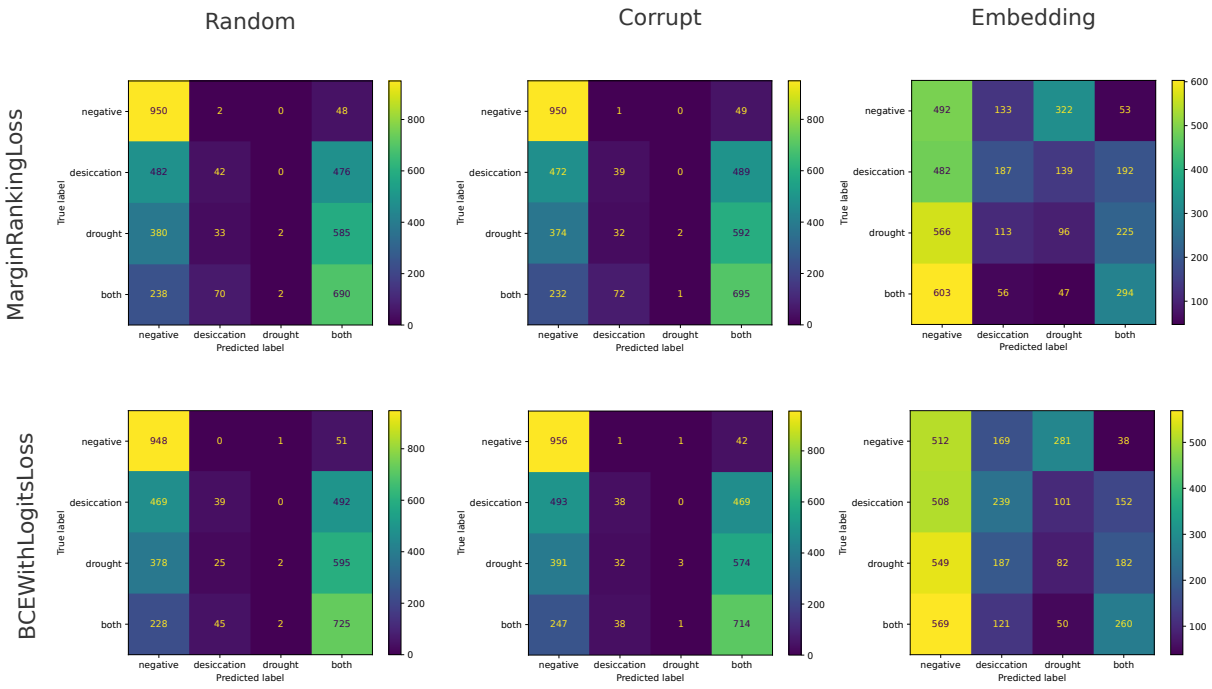


Figure S4.3 **Confusion matrices for RF models with different sampling strategies.** Confusion matrices for models for each upstream RESCAL loss function and negative sampling strategy. No meaningful difference was found between the two RESCAL losses in the RF models, and there was no meaningful difference between the random and corrupt sampling strategies.

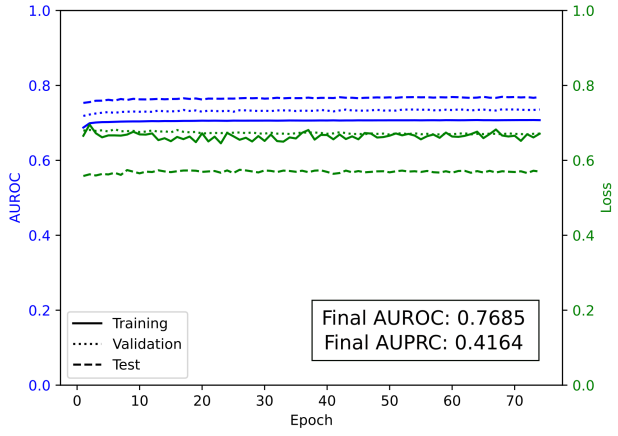


Figure S4.4 **STHN performance.** AUROC and loss plotted for the training epochs of STHN.