Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biology

Marius Dobbe Klemetsen

# Building a Knowledge Network of Regulated Cell Death in *Arabidopsis thaliana*

Master's thesis in Biology: Cell and Molecular Biology
Supervisor: Martin Kuiper
Co-supervisor: Daniela Jorgelina Sueldo, Eirini Tsirvouli
May 2024

NTNU
Norwegian University of
Science and Technology

Marius Dobbe Klemetsen

# Building a Knowledge Network of Regulated Cell Death in *Arabidopsis thaliana*

Master's thesis in Biology: Cell and Molecular Biology
Supervisor: Martin Kuiper
Co-supervisor: Daniela Jorgelina Sueldo, Eirini Tsirvouli
May 2024

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biology

NTNU
Norwegian University of
Science and Technology

# 1 Abstract

Biological systems at the level of gene regulation and molecular interactions are vastly complex. Today, high-throughput molecular methods result in a continuous stream of biological data that subsequently needs to be interpreted to form knowledge. The interpretation of data is time-consuming and requires domain-specific expertise, and this hard-earned knowledge should be utilized to the best extent to formulate descriptive models of the system in question, as well as to formulate new hypotheses to test to get an even deeper understanding of the system. The biological topic of this thesis is the study of the genetically encoded process of cell death in plants. The process of regulated cell death (RCD), central to the context of growth, development, and responses to environmental stressors, is to a large extent still poorly understood and described. The work presented in this thesis aimed to connect the prior knowledge of RCD in the model species *Arabidopsis thaliana* in a so-called knowledge network. This knowledge network will serve as a valuable resource for interpreting new research findings and providing easy access to prior knowledge necessary for developing descriptive conceptual models of plant RCD.

Information about RCD was acquired through a combination of literature review and programmatic retrieval of data from biological databases. The retrieved information included entities (such as genes, proteins, and small molecules) directly described as involved in RCD in *A. thaliana*; entities inferred to be involved through orthology with *Viridiplantae* species or *Homo sapiens*, entities annotated with relevant Gene Ontology Biological Process terms; and entities with experimentally validated interactions with the aforementioned entities. The information was processed in a data pipeline specifically developed for the project. This pipeline also retrieved several relations between the entities, to formulate the resulting knowledge network, consisting of more than 2 000 entities and 9 000 connecting relations. The relations connecting the entities are experimentally validated physical interactions and co-expression of genes, as well as predicted interactions and co-occurrences in the literature.

The knowledge network displays scale-free characteristics and contains amongst others entities that in a study were identified as having upregulated gene expression under conditions known to induce RCD. Consequently, several possible uses of the knowledge network are described in prioritizing entities in research findings according to how likely they are to be part of plant RCD.

# 2 Sammendrag

Regulert celledød (RCD) er en livsviktig prosess som forekommer i både dyr og planter. Til tross viktigheten for både vekst, utvikling og som respons til stressfaktorer i miljøet, er denne prosessen i liten grad beskrevet i planter. Den ervervede kunnskapen om RCD er spredt utover den vitenskapelige litteraturen og mange forskjellige biologiske databaser. For at kunnskapen skal danne grunnlaget for beskrivende modeller for de molekylære systemene, og for å danne nye hypoteser om hvordan de fungerer, må kunnskapen først samles. Dette prosjektet hadde derfor som mål å knytte kunnskapen om RCD, i modellorganismen Arabidopsis thaliana, sammen til et kunnskapsnettverk.

Datagrunnlaget for kunnskapsnettverket er opparbeidet gjennom litteraturstudie og innhenting av data fra biologiske databaser ved bruk av programmeringsmetoder. På dette vis ble informasjon om molekylære enheter og relasjonene de deler sammenfattet.

Det genererte kunnskapsnettverket består av over 2 000 enheter og mer enn 9 000 relasjoner. Kunnskapsnettverket består hovedsakelig av gener og proteiner, som enten allerede er beskrevet som delaktig i RCD, eller kan anses som kandidater på bakgrunn av ortologiske forbindelser eller grunnet fysiske interaksjoner til øvrige enheter i nettverket. Relasjonene mellom enhetene er fysiske interaksjoner, samuttrykk av gener, predikerte interaksjoner, og forekomst i samme artikler.

Kunnskapsnettverket kan brukes til å evaluere og prioritere hvilke eksperimentelle funn er med størst sannsynlighet involvert i plante RCD, og for å tilegne seg kunnskap som kan overføres til modeller for RCD.

# 3 Acknowledgment

This Master's thesis is the culmination of support, guidance, and encouragement from many individuals, to whom I am deeply grateful.

First and foremost, I extend my heartfelt thanks to my supervisors, Martin Kuiper and Daniela Jorgelina Sueldo. Your expertise, guidance, and insights into the workings of academia have been invaluable. You have given me knowledge that goes far beyond the contents of this thesis, for which I am truly grateful.

I also want to thank Eirini Tsirvouli for her precise explanations, which helped me navigate and resolve specific challenges. Åsmund Flobak, your inspiration showed me how the skills acquired through this project can be broadly applied across the field of biology.

I would like to express my gratitude to all the professors and teachers who have significantly contributed to my education and passion for biology. Special thanks to Grzegorz Konert and Laura Jaakola for supervising my Bachelor's thesis, and to Per Kristian Solevåg and Ellen Harris, who first ignited my interest for the study of biology. Thank you also to Bjørn Kiperberg who in the early years sparked my interest for science and nature.

Finally, I am immensely grateful to my peers and family for their unwavering support and encouragement throughout this journey. Your presence and belief in me have been fundamental in reaching this milestone.

# Contents

# 4 Abbreviations

**ACD** Accidental cell death
**AGI** Arabidopsis Genome Initiative
**API** Application Programming Interface
**BP** Biological Process
**CC** Clustering coefficient
**CID** PubChem compound identifier
**DEG** Significantly differently expressed genes
**DOI** Digital Object Identifier
**dPCD** Developmentally induced programmed cell death
**ECO** Evidence and Conclusion Ontology
**EMBL-EBI** European Molecular Biology Laboratory - European Bioinformatics Institute
**ePCD** Environmentally induced programmed cell death
**FAO** Food and Agriculture Organization of the United Nations
**GO** Gene Ontology
**HS** Heat shock
**IBA** GO evidence group: "Biological aspect of ancestor evidence used in manual assertion"
**ID** Identification/identifier
**IEA** GO evidence group: "Evidence used in automatic assertion"
**IMP** GO evidence group: "Mutant phenotype evidence used in manual assertion"
**ISS** GO evidence group: "Sequence similarity evidence used in manual assertion"
**KN** Knowledge network
**KO** Knockout
**MI** Molecular Interaction ontology
**NCBI** National Center for Biotechnology Information
**NCCD** Nomenclature Committee on Cell Death
**ND** GO evidence group: "No evidence data found used in manual assertion"
**OLS** Ordinary Least Square regression
**PAMP** Pathogen-associated molecular patterns
**PCD** Programmed cell death
**PD** Process description
**PMID** PubMed article identifier
**pPCD** Pathogen induced programmed cell death
**PPI** Protein-protein interaction
**RCD** Regulated cell death
**REST API** Representational state transfer application programming interface
**ROS** Reactive oxygen species
**SA** Salicylic acid
**SBGN** Systems Biology Graphical Notation
**SID** PubChem substance identifier
**SKM** Stress Knowledge Map
**TAIR** The Arabidopsis Information Resource
**TF** Transcription factor

# 5 Introduction

## 5.1 Why study cell death in plants?

The Food and Agriculture Organization of the United Nations (FAO) has highlighted in its 2023 annual report that in 2022 almost 30% of the world's population were moderately or severely food insecure. Moreover, in 2021 more than 40% could not afford a healthy diet [1]. These numbers are striking, pinpointing the need for a better understanding of food production, -quality, and -preservation. As photosynthetic organisms are the main primary producers, understanding these organisms should be of high global priority. In addition to food production's impact on human health, land use for agricultural purposes has a high ecological impact. There are also ethical reasons why plant research should be prioritized. The main being that food is a main pillar of life and therefore should be prioritized over research which may only affect the lives of those economically strong. These key points are true as of today, and will likely be of greater importance moving forward with the expected outcomes of global climate change. Increased temperature with pursuing droughts and shifts in the geographical distribution of species will impact the degree of abiotic and biotic stress plants have to endure. If we are to handle the impact these stressors will have, we first need to understand the processes these stressors induce.

A fundamental part of the evolution of life, and the life cycle of an organism, is death, both at the level of whole organisms and at the cellular level. Even so, most articles on plant cell death mention that the research is ongoing and in many areas far behind the research in animals [2]. As cell death is part of the plant's development and in response to certain environmental stressors, it is a fundamental process that needs to be described if we are to fully understand the workings of plants.

Biological processes are complex in that many molecular events form extensive pathways that lead to the outcome if the requirements are met. The research field of plant cell death seems to have described events at both ends of such pathways but lacks an understanding of the intermediary events. Given that molecular pathways overlap and affect one another the intermediary events cannot be overlooked. Although the research in plants is behind the equivalent research in animals there is available data on the topic. The data needs to be organized and interpreted for it to spark new hypotheses. This project is meant to contribute to that process.

## 5.2 Cell death

Cell death plays a crucial role in a plant's life. It is the result of processes occurring at various times during the plant life cycle. Both internal and external factors may induce cell death. Cell death is a necessary means for many critical processes, such as the formation of vascular tissue, the spread of pollen, and the prevention of pathogen proliferation. Although a necessity in certain processes, cell death may also be unwanted and unpreventable. The terms accidental cell death (ACD) and regulated cell death (RCD) can be used to distinguish between processes leading to unwanted death, and death facilitated by the cellular machinery [3, 4].

ACD is the result of stress levels surpassing the physiological tolerance threshold. Examples of instances where this may occur are mechanical injury or extreme abiotic conditions in regard to temperature, pH, salinity, or radiation [5]. Under such conditions death to the cells are inevitable. Under normal physiological conditions, cell death may occur, then as the result of RCD. The cellular machinery that facilitates RCD is genetically encoded. As genetically encoded entities may be regulated at many levels, it follows that the processes they are involved in may also be regulated. The property of regulation means that RCD can be inhibited or at least adjusted, i.e. death is not inevitable.

In this project, I will keep to the term RCD to describe all processes leading to cell death in a regulated fashion. The term programmed cell death (PCD) has historically been, and often continues to be, used as an umbrella

term for all cell death processes that are not ACD. However, for animal systems the Nomenclature Committee on Cell Death has suggested the use of the term RCD [5], as well as many authors of articles on cell death in plants [6]. The term PCD may, however, be more suited to describe cell death processes occurring under physiological conditions. PCD has been used in this sense for the further distinction between PCD induced as part of development (dPCD) and PCD induced by environmentally derived cues (ePCD) [7]. As environmentally derived cues can be abiotic or biologically derived, another distinction should be in order. The term pathogen-triggered PCD (pPCD) has been used in the context of RCD induced by pathogens [8], which may suit well as a subcategory of ePCD.

Given that animals and plants belong to two different kingdoms of life one should be careful when extrapolating terms and knowledge that is known from animals to plants. Extrapolation should be reserved for entities or processes that are conserved from common ancestry. Identifying such cases can be challenging, as similarities between entities or processes in different taxonomic clades may stem from convergent evolution. Historically extrapolation of terms has been an issue as molecular knowledge in animal systems has been described earlier or in greater detail. The result has been the use of terms that end with "-like", e.g. apoptotic-like and caspase-like, to describe processes or molecular players in RCD processes in plants. Given that neither apoptosis nor caspases are present in plants [9, 10], using these terms may only cause confusion.

## 5.3 Network science

### 5.3.1 Networks and models

**Networks**

A network is a system comprised of entities and the relations that are found between them. What the entities and relations represent depends on the system in question. Nature is complex, but can be explained as an overwhelming collection of networks. Biological networks range from the higher-order networks of species' ecological roles in the ecosystem to molecular interactions within a single cell. Even at the molecular level, the complexity is almost impossible to comprehend. Biological entities such as genes, RNA, or proteins interact and give expected results, but the vast variety of molecular entities and the combinations of how they affect each other and their environment give rise to incredible complexity. We can try to understand nature by recreating these real-world networks. This task should be feasible but will require an extensive quantity of data on the biological system we want to describe. Today, this data is rapidly being collected through high-throughput laboratory methods however, for the data to give rise to the desired models it first has to be interpreted, which often requires deep knowledge of the topic.

**Dynamical models, maps, and knowledge networks**

Dynamical models are suitable for simulating real-world systems and must be considered the ultimate goal of computational biology. The quality of such models may vary, but all would require a great deal of knowledge to be generated, moreover for them to give reasonable results. Such models require all types of parameter values, like enzyme kinetics, gene transcriptional requirements, gene transcriptional rates, environmental composition, etc. Dynamical models are desirable as simulations can imply the result of perturbation or variable input to the system.

A map is a conceptual model of a mechanistic system. Maps are visual representations of the system in question, that are highly accurate, and arguably the best format to present the system to an audience. As with dynamical models, a great amount of knowledge is required to build informative maps. However, the requirements may often be lower, resulting in maps often being predecessors of dynamic models [11]. The Systems Biology Graphical Notation (SBGN) is a graphical format that was formulated to be able to visually present biological systems with high accuracy. The system is meant to present the workings of the entities in an unambiguous and more structured way than what can be described with text. The SBGN has three formats, Activity Flow,

Entity Relationship, and Process Description, with each presenting a greater amount of information [12]. The objective of this project is to generate a resource that can help in the process of generating maps.

The focus of this project will be to generate a so-called knowledge network (KN). A KN is a network that includes a broader range of relation types than the causal interactions often seen in dynamic models and maps. The relations can be both physical and abstract. For instance, the KN produced in this project contains physical interactions between proteins, predicted relations, and relations found between the text representation of the proteins. In this sense, KN can simply be described as a collection of prior knowledge, represented in the format of a network. A KN is meant to be a stepping stone toward dynamical models and maps, in that it can contain knowledge that can be directly transferred to them or knowledge that can spark hypothesis generation.

**Networks and graph theory**

The benefit of a network representation of knowledge is that a network can be viewed as a mathematical object with mathematical properties, as described by graph theory. Although in mathematics a network is often referred to as a graph I will for continuity refer to it as a network. In mathematical terms, the entities of a network are known as nodes or vertices, while the relations between the entities are known as links or edges. As a mathematical object, a network also has mathematical properties. As will be described further, these properties can showcase information that would be difficult to observe had the data not been formulated as a network.

A node's degree, i.e. the number of links that connect a node to other nodes in the network, is an important parameter used to assess the node's role in the network. Highly connected nodes are known as hubs. A node's role in the network also depends on the type of relation links it shares with other nodes. The distribution of all the node degrees infers certain characteristics of the network. For instance, with a power-law distribution, the network is characterized as a scale-free network. In such networks, the largest fraction of nodes will have a low degree, while a considerable small fraction will have a very high degree. This distribution is the result of hub formations in the network, which is highly characteristic of biological networks.

The properties of a node also depend on the nodes that it is directly connected to, i.e. its neighborhood. The influence depends on what the nodes of the neighborhood represent, and the type of link they are connected by. Moreover, how the nodes in the neighborhood are interconnected is also important for the given node. The clustering coefficient (CC) of a node describes exactly that. The CC value ranges from 0 to 1. If all the nodes in the neighborhood are interconnected the CC is 1, while 0 if none are interconnected. Groups of nodes that are fully interconnected are known as cliques. A node with a CC value of 1 will always be part of a clique. However, members of a clique can have other CC values if they are also connected to nodes outside of the clique. A biological example of a clique can be proteins interacting to form a protein complex. Generally, groups of nodes that are more connected within than to the outside of the group are said to form a cluster (a.k.a. a community).

## 5.4 Ontology, orthology, and relations

In this project, entities are retrieved and incorporated into the knowledge network on the basis of annotations by Gene Ontology terms and on the basis of orthology, and will therefore be described here in brief. The same applies to the relations found in the knowledge network.

### 5.4.1 Ontologies, the Gene Ontology, and Controlled Vocabulary

An ontology is a hierarchical structure consisting of classes, where each unambiguous class represents a specific feature. Classes are organized in a hierarchical format so that the uppermost classes represent broad features that can be split into more specific features. In that sense, a more specific class shares a relationship with a

broader class if the feature itself represents is also part of the feature the broader class represents. In ontologies, there might be relationships between classes in directions other than the vertical axis (i.e. between parent and child class), although a description of how this works is not necessary to understand this project and will therefore not be given here. Ontologies can be applied to all types of systems to describe the entities of the system with as much accuracy as possible while maintaining unambiguity in the description [13].

The Gene Ontology (geneontology.org) (GO) is probably the most utilized ontology in biology. It is an ontology that describes the features of a gene and its gene products (i.e. RNA or proteins). In the GO, a class is known as a term, and a higher-order term can be split into child terms, which then share a relationship with the parent term. A gene or gene product that has the feature that the GO term represents is said to be annotated with the given GO term. At the uppermost level the GO is divided into the three root terms; Molecular Function, Cellular Component, and Biological Process (BP), all describing different aspects of the gene or gene products. All terms in the GO are thereby descendant terms of either one or more of these root terms. A simplified explanation of the Molecular Function, Cellular Component, and BF GO terms is that they respectively describe what the gene product does, where it is located, and when it contributes to a process. More thoroughly, a Molecular Function GO term describes what the annotated gene product is capable of at a molecular level, which often means describing the activity of the gene product in a reaction. A Cellular Component GO term describes the physical location of the gene product within the cell, e.g. plasma membrane, cytosol, or other cellular compartments. The BF GO term describes which biological processes the gene product is involved in. It is important to note that the individual gene products of a gene can be annotated with different GO terms, as they can have different properties, however, the encoding gene will be annotated with the combined total of all the gene products' GO term annotations. The GO is supposed to be as inclusive as possible, meaning that it should be applicable for annotating any gene and gene product from any species. An example of a subsection of the GO is seen in figure 1 [14, 15].

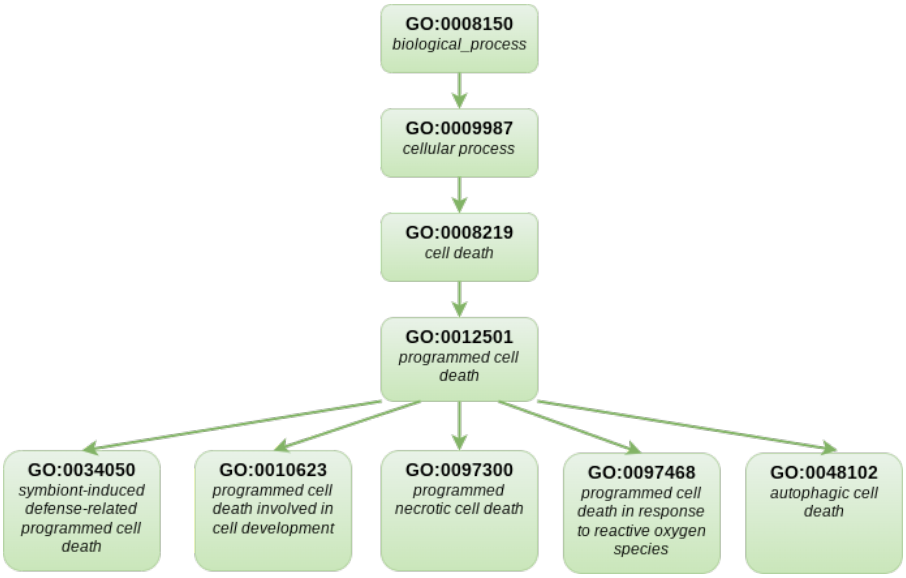



*Figure 1: An example of how the GO Biological Process term "programmed cell death" (GO:0012501) has relations to higher-order GO terms (i.e. ancestral terms) and lower-order GO terms (i.e. descendant terms). This example shows only the first descendant terms (i.e. the child terms) currently used for gene and gene products in the taxonomic clade Viridiplantae (i.e. plants).*

In the same way that the GO describes features of genes and gene products, there is a separate ontology that describes the evidential grounds for why a gene or gene product was annotated with a GO term. This ontology is known as the Evidence and Conclusion Ontology (ECO) [16]. As with the GO, the ECO contains unambiguous terms, hereafter referred to as evidence codes. A high-level organization groups the evidence codes into; "Experimental evidence codes", "Phylogenetically-inferred annotations", "Computational analysis evidence codes", "Author statement evidence codes", "Curator statement evidence codes", and "Electronic

annotation evidence code" [15]. The latter group contains only the single evidence code of "Inferred from Electronic Annotation" (IEA). The IEA code is the most prominent evidence code for all the annotations (in the QuickGO database), with almost 99.6% of annotations having this or descendant codes (in the GO version 2024-05-01).

A controlled vocabulary that is referred to in this project is the Molecular Interactions Controlled Vocabulary (PSI-MI). This controlled vocabulary is used to describe molecular interactions as well as the methods used to detect these interactions [17]. In this project, references to terms of this controlled vocabulary (version: 2.5.5) are given with the prefix "MI:" followed by a four-digit number.

### 5.4.2 Orthologous genes

Two similar genes are described as homologous if they stem from a common ancestral gene. During speciation events, many of the same genes will be retained in the genome of the respective species, where they can evolve independently. Genes that originate from an ancestral gene and are retained in the genome of the respective species following a speciation event are known as orthologous genes (or orthologs). Genes within the genome of a species can be duplicated, with the two copies being described as paralogous genes (or paralogs). The evolutionary relationship between genes can be described further through the terms co-orthologs, out-paralogs, and in-paralogs, although this will not be necessary to understand this project [18].

It is important to discover and study orthologs because the biological function of the ortholog in one species may be similar in the other species. This means that information acquired for one species may also apply to another species. This latter remark is quite obvious for evolutionary closely related species, as they are expected to have a higher degree of genetic similarity than those more distantly related. When combining the knowledge of orthology and functional aspects of the genes and gene products in question we can hypothesize aspects of other genes across species. To what extent expect conserved functional aspects will depend on the evolutionary relationship. In this project orthology between human and *A. thaliana* proteins are assessed which of course are quite distantly related. It is important to remember that although the biological function may be similar between orthologs, many other aspects determine if the role of a gene is the same as the orthologous gene in their respective species. For instance, gene regulation will extensively affect the gene's role, in that the gene may be expressed in different quantities or at different periods if it even is expressed.

### 5.4.3 Interaction events and other relations between entities

In this project, I will use the term interaction when describing a physical interaction occurring between two entities, and the term relation when there is a relationship between two entities that is not a direct physical interaction. Examples of the latter are a co-expression relation or co-occurrence relation in text.

**Physical interactions**

Physical interactions here refer to an interaction occurring between two entities because they are in physical proximity to each other. This definition is very broad and includes colocalization and molecular association, which are described respectively by the Molecular Interactions Controlled Vocabulary codes MI:0403 and MI:2232, or descendant codes. PPIs will fall within this category. Typical PPIs are complex formation and regulatory interaction events such as phosphorylation and dephosphorylation. In the knowledge network, all physical interactions can be described by descendant terms of direct interaction (MI:0407) are annotated as "special" interactions.

**Co-expressed genes**

Genes that are expressed in similar ratios for a given condition are described as being co-expressed. A co-expression relationship can mean that the co-expressed genes are transcriptionally regulated in similar fashions,

i.e. they may share transcription factors.

**Co-occurrence relations identified with text mining methods**
Abstracts of scientific articles tend to precisely and compactly describe the content of the article. These features make abstracts suitable for extracting information on potential relationships between genes, gene products, and biological processes. As abstracts are more readily available than full articles these are often used in automated processes of data retrieval. Such automated processes extracting information from human-written text are known as text mining processes. The co-occurrence relations identified by STRING for the most part observed in article abstracts, although some may be co-occurrences in articles that are freely available. STRING calculates a co-occurrence score based on how close in the text the two entities are mentioned [19, 20].

## 5.5 Biological databases and network software

The status quo of research is that research findings are published as articles in scientific journals for the community to read. However, currently, there are extensive quantities of articles being published for researchers to keep track of. Domain-specific databases that contain knowledge extracted from scientific articles can massively increase the accessibility of research findings. A database entry is of great value if it has been annotated thoroughly. Annotations may have been contributed by experts, although many databases also utilize programs for automatic annotation. For increased accessibility, these databases should provide Application Programming Interfaces (APIs) making the data available by programmatic means. By providing APIs the databases facilitate the possibilities of large projects with retrieval of data on many biological entities. Moreover, APIs provide a more sustainable method of incorporating new data into projects whenever the content of the database is updated.

Structured databases are of great use to assess knowledge and to easily request changes or highlight knowledge gaps. There is a plethora of biological databases with varying degrees of quality, coverage, and maintenance. However, there are a lot of redundancies in the information presented in the individual databases which can be beneficial when tracking e.g. a specific gene, although it can present a challenge to identify what sets the given database apart. There are efforts to coordinate the development of biological databases like the European initiative ELIXIR (elixir-europe.org) (part of European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI)). All the databases described in the following section are part of the ELIXIR initiative, except GeneMANIA, TAIR, and PubChem.

### 5.5.1 Biological databases - Storage of knowledge

**UniProt**
UniProt (uniprot.org) has been the main source of data in this project. UniProt is dedicated to storing protein and proteomic data but in many ways also serves as a hub for knowledge retrieval thanks to the comprehensive cross-referencing to other databases. One of the benefits of UniProt is that each protein isoform encoded by a gene is stored as separate entries with unambiguous stable identifiers, thereby uniquely identifying the gene data from the protein data. These unique identifiers, known as UniProt accession codes, are widely searchable in other databases, especially in ELIXIR databases. In practice, UniProt consists of two different databases Swiss-Prot and TrEMBL, which respectively store entries that have been expertly reviewed and manually annotated or entries that have been automatically curated and annotated [21].

**IntAct**
IntAct (ebi.ac.uk/intact) stores data on experimentally validated physical interactions between proteins (i.e. protein-protein interactions (PPIs)) or proteins and other biological entities (e.g. DNA). Each interaction has its unique accession code and annotations describing the interacting entities, the detection method, the host organism, and the publication where this interaction was identified (often presented as the PubMed reference

code (i.e. PMID)). The two, or more, interacting entities (known as the interactors) are identified with identifiers, which for PPIs is most often the UniProt accession code. The host organism is in what system the interaction was observed, being either an organism (e.g. *Saccharomyces cerevisiae* (Baker's yeast) for a yeast two-hybrid assay) or *in vitro*. IntAct provides an evidence score of the interaction known as the MIscore. The MIscore is calculated based on the detection method, the interaction type, and the number of publications identifying the interaction [22]. The MIscore is useful when trying to determine which interactions need further experimental tests to be properly validated. IntAct does not provide the context of the interaction (e.g. environmental conditions), although this will likely be disclosed in the original paper(s) [23].

**QuickGO**

In this project, the GO annotations were accessed through the QuickGO (ebi.ac.uk/QuickGO) database. QuickGO facilitates easy search for a given GO term and the resulting annotations, i.e. the genes annotated with the given GO term. The search options are tunable, however, by default, QuickGO returns genes annotated with the inputted GO term or descendants (also known as child terms) of this term. The same is true regarding the inputted taxonomic level (i.e. the taxonomic identifier), meaning that all lower-level taxons are included (e.g. if primates were inputted, all species of primates would be searched). I.e. a higher-order search for cellular component of the mitochondrion, in all vascular plants (i.e. *Tracheophyta*), will return all gene products in any vascular plant that are located in any localization within the mitochondria [24].

**OrthoDB**

The database called OrthoDB (orthodb.org) stores an overview of which genes can be considered as orthologous. The orthologs are identified by comparing reference genomes programmatically. The benefit of such a method is that orthologs may be identified without any prior knowledge of the genome other than sequence. As genes that stem from the same ancestral gene tend to have conserved functions, annotations of one ortholog may be transferable to another. OrthoDB organizes all genes identified as orthologs into respective groups, also known as clusters. These ortholog groups are arranged according to the taxonomic hierarchy, meaning that orthologs within a group are genes from species within the same taxonomic clade. In practice, the ortholog groups in OrthoDB contain the individual gene products, identified by their UniProt accession codes, rather than the genes [25].

**InterPro**

InterPro (ebi.ac.uk/interpro) stores information on the classification of protein families, domains, and functional sites. This database is a collective resource incorporating several protein annotation databases, such as Pfam and PANTHER. Domains and functional sites are crucial to the functional properties of a protein. Identifying specific domains or functional sites in proteins that are potential components of certain pathways can provide stronger evidence that they are part of those pathways [26].

**STRING**

STRING (string-db.org) is a database that stores protein information. The information is collected from curated databases to form a large repository of knowledge. More importantly, STRING connects the proteins in networks. The relations found between the proteins can be experimentally validated relations (e.g. PPIs and co-expression relations) or other types of relations. One of these other relations is the co-occurrence of proteins in PubMed abstracts. The co-occurrences are identified using automated text-mining procedures [20].

**GeneMANIA**

The GeneMANIA (genemania.org) database performs a similar function as STRING by retrieving data from numerous other databases and datasets, organizing it, and connecting it. GeneMANIA returns relations between gene and gene products identified with experimental data, like physical interactions and co-expression, in addition to predicted relations often stemming from PPIs of orthologs [27].

**TAIR**

The Arabidopsis Information Resource (TAIR) (arabidopsis.org) is a database that in many cases mimics the functionality of UniProt but is dedicated specifically to storing information on the model plant *A. thaliana*. It is managed by Phoenix Bioinformatics Corporation. The major drawback with TAIR is the paywall and the lack of a user-friendly API. Much of the information provided by TAIR can, however, be accessed through other databases. In this project, TAIR was mainly used in cases where certain information regarding a gene product was not present at UniProt [28].

**Ensembl Plants**

Ensembl Plants (plants.ensembl.org) is the plant domain-specific part of the Ensembl Genomes. Ensembl Plants provides genome annotations [29]. Ensembl Plants also provides the readily used tool called BioMart (plants.ensembl.org/biomart/martview). BioMart is an identifier mapping tool, meaning that it retrieves all identifiers linking to the same biological entity. In this project, BioMart was utilized to get the UniProt accession code(s) from the various identifiers returned by the other utilized resources.

**PubChem**

PubChem (pubchem.ncbi.nlm.nih.gov) is a database dedicated to the storage of a variety of chemical data. What sets PubChem apart from the other mentioned databases, is the storage of information on other chemical substances than just genes or gene products. PubChem is part of the American National Center for Biotechnology Information (NCBI) database collective [30]. In this project, PubChem was utilized to retrieve annotation data for chemical substances not listed in UniProt, i.e. ions and small molecules.

### 5.5.2 Network software - generating and presenting networks

There are many software tools available for the generation of graphical networks. Some serve a general purpose while others are specialized for biological networks. For this project, the utilized software were those that allowed for bulk import of all the data of the network. Software requiring a lot of manual intervention for network generation is not suitable when the network consists of thousands of nodes and links. Given the large size of the network in this project, the software was chosen accordingly.

The software yEd (distributed by yWorks) is an example of a general network software that was utilized in this project [31]. yEd has the option of importing the network data as a Microsoft Excel file wherein the nodes and their annotations are separated from the links and their annotations. yEd was deemed to have a user-friendly interface that makes it easier for the user to observe interesting aspects of the network and annotations to the nodes and links.

Cytoscape is the other network software utilized in this project. It was developed with biological networks in mind and contains a lot of features that can be of great use in computational biology projects. Moreover, it houses a store of externally developed applications that can enhance the usefulness of Cytoscape [32]. For this project, Cytoscape was mainly used for the feature that utilizes graph theory to calculate node properties.

Large networks are notoriously difficult to present in a static format that effectively conveys information. Therefore, it is important to remember that a network is often more informative when interacted with. Interpreting smaller sections or individual nodes helps to identify interesting aspects and understand their connections to other parts of the network.

## 5.6 Data pipelines and application programming interfaces

A data pipeline is a term that describes a process where data is inputted by the user, processed, and then returned in a defined format. In this project, some data is inputted by the user, while some are imported from

databases using their APIs. The returned output is the knowledge network. A pipeline is meant to be reusable, which this project's pipeline is, meaning that the user input can be changed to generate alternative output in the same format. The data pipeline produced in this project is generated using the highly utilized programming language called Python. Using a programming language with a high user base is beneficial in ensuring further development of the data pipeline.

## 5.7 Objectives and approaches

The first objective is to assess the status of gene and gene product annotation of cell death-related GO BP terms in the taxonomic clade *Viridiplantae* and more specifically in *A. thaliana*. The knowledge gathered from this assessment will be the basis for how the KN is generated. Specifically, it will determine which GO BP term annotations will be accepted on the basis of the evidence of their annotation.

The second and main objective of this project is to generate a knowledge network (KN), containing both validated and putative entities and relations involved in RCD processes in *Arabidopsis thaliana*. Experimentalists can subsequently use this KN during the analysis of their experimental data, check their gene lists against, and to generate hypotheses that can be tested in the lab, to enhance our understanding of RCD in plants. The KN will be based on information gathered by manual and automatic means. Moreover, it will be largely based on protein annotations of cell death-related GO BP terms, orthology, experimentally validated interaction events, co-expression, and co-occurrence in article abstracts. The KN will structurally display what information can be retrieved from biological databases, how the information was discovered, and what the quality of evidence is that it is supported by. Putative entities and interactions must be described with reasoning for the inclusion, and show a traceable line going back to the original reference (provenance). With this in place, a user has the option of trusting the KN fabricators or easily accessing the source material to judge for themselves. Additional information should be easily available to the user, to reduce the time resources spent finding information that can be rapidly acquired using automatic processes.

The third objective is to present use cases of the KN. The use cases will be based on interesting network properties as well as how the KN can be utilized in combination with external experimental data from a study on RCD in *A. thaliana*.

The fourth objective of this project is to formulate methods incorporated as part of a data pipeline that facilitates the generation of new KN. A KN should be as relevant as possible meaning that the continuously updated knowledge should be incorporated into the KN. Moreover, the KN network generated in this project is the result of my decisions and reasonings, which may differ from others. The objective in formulating the data pipeline therefore is for it to be easily adjustable and give the option of receiving data presented by the user, here being manually curated entities and relations.

# 6 Materials and methods

In this project, two methods to retrieve entities and relations for the knowledge network (KN) were used. The first method was manually curating information from the scientific literature, while the second was retrieval of information from curated databases by programmatic means. A schematic overview of the materials and methods used is shown in figure 2. The data regarding the nodes (entities) and links (relations) of the network were stored in respective tables. This format made for easy distinctions of what annotations belonged to the nodes and what belonged to the links.

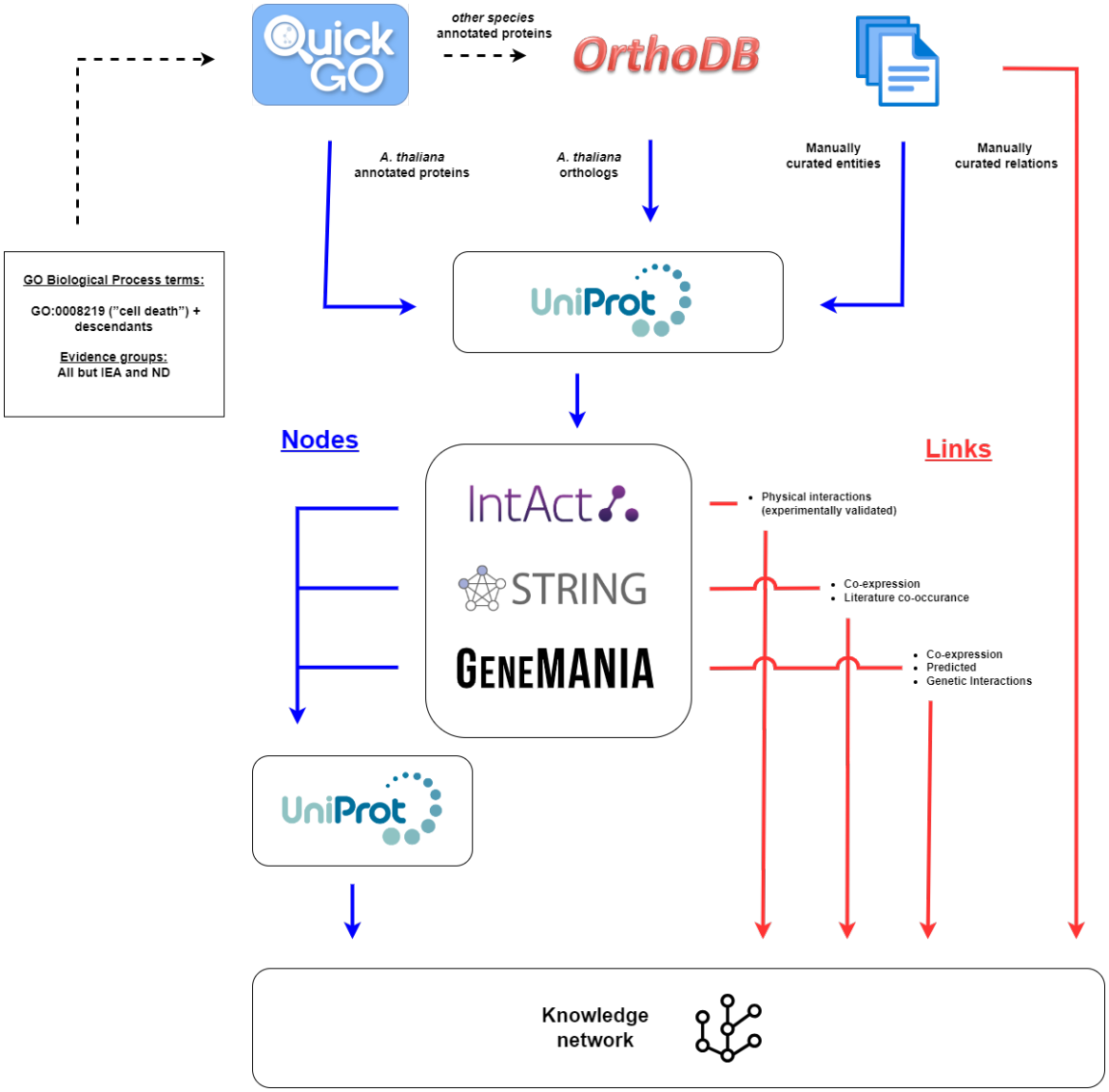



*Figure 2: Schematic overview of the methods and how the data from the different databases were retrieved sequentially until the final inclusion in the knowledge network. The databases are indicated with their logos. Dashed lines indicate what was searched for in the database pointed to, while solid lines indicate processes where data is retrieved. Blue arrows indicate processes related to the retrieval of data for the entities (i.e. the nodes), while red arrows indicate processes related to the retrieval of data for the relations (i.e. the links).*

## 6.1 Manual data curation

The criteria used for the manual curation of entities and relations were broadly inclusive. From the literature, any entity or relation that seemed to impact or be involved in RCD in plants was curated. The entities and relations curated had to be connected to the plant model species *Arabidopsis thaliana*. The connection could be that the entities were derived from or naturally occurring in *A. thaliana* cells, e.g. genes, proteins, hormones, metabolites, ions, or other molecules shown to impact RCD in *A. thaliana* cells. These latter molecules were

either produced by other species (e.g. pathogen-derived) or were not biologically derived (e.g. Lanthanum(III) chloride). Data retrieved for the manually curated entities included molecular identifiers (the UniProt accession code or PubChem compound ID respectively for proteins or other chemical substances), as well as the reference where I first identified the entity. The preferred reference was the PubMed article entry ID (PMID), while the digital object identifier (DOI) code was noted in cases where the PMID could not be found. Similarly, the relations were annotated with the identifiers of the partners sharing the relation (i.e. the interactors), the literature reference, the type of literature (e.g. article or book chapter), and the type of interaction (e.g. physical interaction or catalytic participation in a reaction). In addition, some notes regarding the subjective reasoning for curation were annotated to both the curated entities and relations. The literature used was mainly review articles and books on the topic of plant RCD (i.e. secondary sources).

## 6.2 Automatic data retrieval

The following sections describe how data were acquired for the knowledge network (KN) through programmatic means using the APIs of the respective databases and the programming language Python (version: 3.11.4). The following section is for simplicity described by sequential steps, although there might be some overlap between the processes. The exact sequential steps can be inferred from the Python code. A schematic overview of the process is given as figure 2. The exact values and settings given in the following section are what was chosen for the generation of the KN presented in the result section. Although another user may choose different settings, the sequential steps will be the same.

### 6.2.1 Fetching proteins annotated with GO terms relevant to RCD

From the QuickGO database, all proteins annotated with the GO BP term "cell death" (GO:0008219) or descendants of this term (hereafter referred to as cell death-related GO BP terms), in the clade *Viridiplantae* (taxonomic ID: 33090) and *Homo sapiens* (human) (taxonomic ID: 9606), were acquired. These results were stored in a separate file and used to present as part of the results to give reason for why proteins with annotations with certain evidence groups were not retrieved and included in the knowledge network, moreover used to retrieve *A. thaliana* orthologs.

With this rationale, which will be discussed later, the results were filtered to not include proteins annotated by automatic means, i.e. those with evidence code of "Evidence used in automatic assertion" (IEA) (ECO:0000501), or descendants of this evidence term. For *Viridiplantae* this meant excluding 97% (15 770 total, 410 filtered) of the annotated proteins, while for human proteins about 75% (2 039 total, 519 filtered). The full query details are shown in Appendix A.1. Similarly, annotations with the evidence code "No evidence data found used in manual assertion" (ND) (ECO:0000307) were excluded, as annotations with this evidence code have no experimental data that indicate that the entity may contribute to the biological process. Only the protein entries from QuickGO with valid UniProt accession codes were included in the KN.

### 6.2.2 Fetching *A. thaliana* orthologs of RCD proteins in other species

The proteins of human and other *Viridiplantae* species manually annotated by experts with cell death-related GO BP terms were used to look for protein orthologs in *A. thaliana*. The UniProt accession codes retrieved from QuickGO were used to retrieve the OrthoDB (version: v11) cross-reference from the UniProt entry page of the given accession code. The OrthoDB cross-reference is the OrthoDB cluster code. The cluster contains all orthologous genes found between reference genomes of those organisms inspected by OrthoDB. Each orthologous gene is presented in OrthoDB with the UniProt accession codes of gene products. If a given cluster contained *A. thaliana* UniProt accession codes, these codes were retrieved. In a separate file was stored the orthologous relation between the protein(s) of the *other* species and the protein(s) of *A. thaliana*. The *A. thaliana* orthologs were included in the KN as putative entities of RCD in *A. thaliana*.

### 6.2.3 Fetching experimentally validated interactions between entities in the KN

All the UniProt accession codes for the proteins included in the KN until this point were used to search for interacting partners in the IntAct database (version: 1.0.4). The codes were used in a batch search using IntAct's API. The search result is experimentally validated physical interactions between entities. The interactions can be between the entities searched for and other entities also having interactions with the former. In this project, only protein-protein interactions (PPIs) were retained. This means that non-proteins were removed from the search results. The full search query settings and how the API output results were handled are given in Appendix A.2.

### 6.2.4 Fetching other relations between proteins in the KN

Relations other than physical interactions between proteins in the KN were fetched from the STRING database (version: 11.5), using its API, and from GeneMANIA, using its web application. From STRING, relations of co-expressed genes and co-occurrence in article abstracts were retrieved. From GeneMANIA were retrieved co-expression relations, predicted interactions, and genetic interactions.

The STRING database API takes UniProt accession codes as valid input but will convert these to the STRING identifiers as part of the returned data. The format of the STRING identifier includes the gene locus code, which was extracted for all the returned entities. Similarly, GeneMANIA takes UniProt accession codes as valid input but returns different types of identifiers. The loaded UniProt accession codes are returned unchanged, but new proteins are returned with the GeneMANIA in-house identifier and the identifier used in the NCBI Gene database. To make sure that all proteins in the KN were identifiable by UniProt accessions, the resulting gene locus codes and NCBI Gene IDs were converted to the corresponding UniProt accession codes using BioMart (version 0.7) from Ensembl Plants. As BioMart tended to return more than one identifier per loaded gene, some filtering was performed to get the most suitable identifier per gene. The most suitable UniProt identifiers were deemed to be the ones linking to the Swiss-Prot part of UniProt. A full description of the BioMart settings and filtering rules for the returned results is given in Appendix B.1.

### 6.2.5 Fetching data on the entities of the KN

Most of the annotation data for the entities of the KN were retrieved by programmatic means. For the proteins, the annotation data were acquired from UniProt using the API with the UniProt accession codes. An example of retrieved data can be seen in table 1. Some additional information regarding the InterPro (version: 99.0) annotations was fetched separately using InterPro's API. For other chemical substances, annotation data were retrieved from PubChem using its API and the compound (CID) or substance (SID) identifiers. The programmatically acquired annotation data were incorporated, along with annotation data acquired through earlier described methods (e.g. curator notes, how the entity was identified, and potential annotations of cell death-related GO BP terms), entity data file.

*Table 1: Example of annotation data retrieved from UniProt entries. The example is an arbitrarily chosen UniProt entry. Multiple annotations per data type were separated by "|".*

| Data type | Example data from UniProt entry Q9SCU7 |
|---|---|
| Protein name | Transcription factor MYB30 |
| Gene name | MYB30 |
| Gene synonyms | hsr1 |
| Species name | Arabidopsis thaliana |
| TAIR identifier | AT3G28910 |
| OrthoDB cluster | 887435at2759 |
| Cell death-related GO BP annotation | GO:0009626 (IEP) |
| InterPro annotations | IPR009057 \| IPR017930 \| IPR001005 |

## 6.3 Methods of handling the data and performing network analysis

### 6.3.1 Handling data

The data was for the most part handled using Python programming (version 3.11.4), specifically with the use of the Pandas library (version 2.0.3) [33]. The KN, or subsections of it, as displayed in the results section were graphically generated using the yEd graph editor (version 3.23.2) (distributed by yWorks) [31].

### 6.3.2 Network analysis

The calculation of network metrics was performed in Cytoscape (version 3.10.1) [32] with the core app NetworkAnalyzer (version 4.5.0). The diagrams of the result section were generated using the Python libraries matplotlib (version 3.7.1) [34] and Plotly (version 5.15.0) [35].

# 7 Results

This project aimed to generate a knowledge network (KN) for the process of plant RCD in *A. thaliana*. The KN should contain data from multiple sources, and display them as a collective in a fashion that makes it possible to generate new hypotheses for how RCD occurs. Moreover, it was meant to be an aid for experimentalists, to put their experimental results in a broader context.

The KN consists of putative and validated entities involved in plant RCD in *A. thaliana*, and the relations they have with each other. The entities include *A. thaliana* proteins annotated with cell death-related GO BP terms, *A. thaliana* protein orthologs of proteins that are annotated with cell death-related GO BP terms in other *Viridiplantae* species, or *Homo sapiens*, biological or non-biological entities that were mentioned in the literature as involved in RCD in *A. thaliana*, or entities that have been experimentally tested to interact with any of the previously mentioned. The relations between the entities are physical interactions, co-expression, co-occurrence in article abstracts, predicted interactions, and genetic interactions.

The methods to acquire the necessary data had to be formulated to generate the KN. The methods are incorporated as part of a data pipeline that accepts a combination of user-provided data and settings, used to retrieve more data from biological databases, to return the collected data of the KN. For the data pipeline to generate a KN with the wanted qualities, some exploration of the input data and settings had to be performed. The following sections will therefore also show why only certain annotation evidence was accepted for the GO term annotations used for the generation of the KN of this project.

## 7.1 Current state of cell death-related GO annotations in *Viridiplantae*

The state of GO annotations of gene and gene products of an organism can tell a lot about how well functional aspects of the organism have been investigated. This project hinges on the use of GO annotations and the following section will therefore highlight the most important elements of the state of GO annotations to proteins of species in the clade *Viridiplantae*. The following results were used as reasoning for why not to include proteins annotated with certain evidence in the KN.

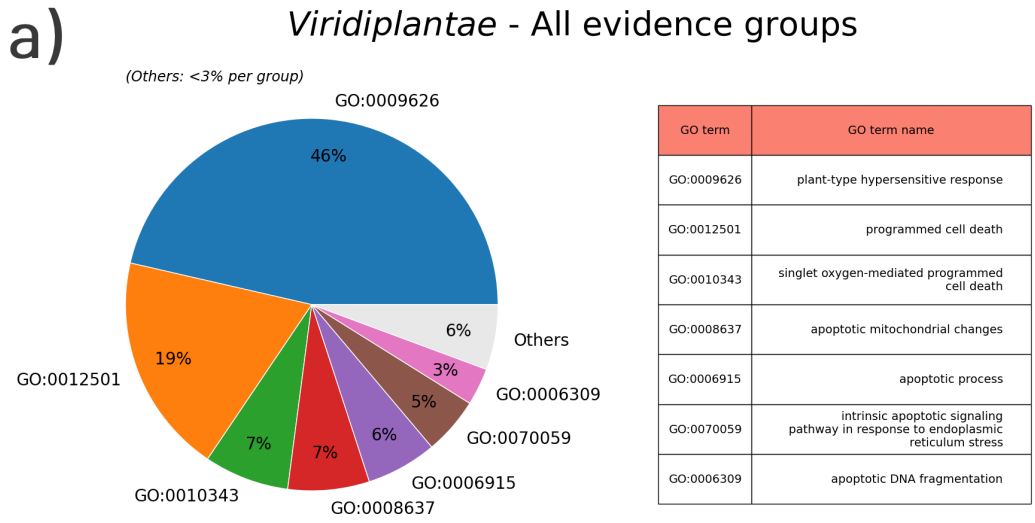
**Cell death-related GO BP annotation of proteins of *Viridiplantae* species**

With the GO version used in this project (version 2024-05-01), 15 770 proteins from 467 species in the clade *Viridiplantae* (which includes *A. thaliana*) were found to be annotated with one or more cell death-related GO BP terms (i.e. annotated with "cell death" (GO:0008219) or descendants of this term). As a protein can be annotated with multiple terms, the number of unique annotations was 19 317. The GO term for the individual 19 317 annotations was found to be one of 25 cell death-related GO BP terms (all terms and their frequencies are listed in table 6 found in Appendix C.1). Most annotations had the term "plant-type hypersensitive response" (GO:0009626) (46%, 8 965 total) or "programmed cell death" (GO:0012501) (19%, 3 698 total). The frequencies of the most prevalent terms are displayed in figure 3a. Almost all annotations (98%, 18 880 total) had the evidence of group IEA ("evidence used in automatic assertion"), indicating that they were the result of automatic annotation. Another remark is that many of the annotated terms contain the word "apoptosis", which is a term that should be avoided when describing processes of plant RCD [9, 10]. For instance, 1 185 annotations had the term "apoptotic process" (GO:0006915).

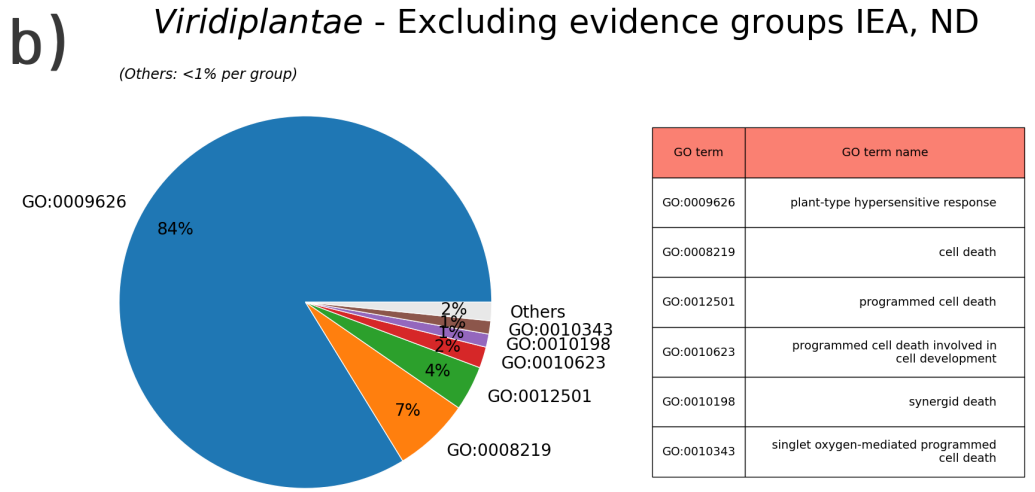
When the annotations with the evidence group of IEA and ND ("no evidence data found used in manual assertion") are ignored, only 437 annotations of *Viridiplantae* proteins remain. These annotations are manual annotations assigned by experts. The annotations have one of 11 GO terms (all terms and their frequencies are listed in table 7 found in Appendix C.1). The frequencies of the most prevalent terms are displayed in figure 3b. The most prevalent GO term (84%, 366 total) is "plant-type hypersensitive response" (GO:0009626). The

annotations are to proteins of 35 *Viridiplantae* species, where *A. thaliana* has the largest number of annotations (29%, 127 total). Most of the annotations (68%, 297 total) have been annotated with evidence in the evidence group IBA ("biological aspect of ancestor evidence used in manual assertion"), meaning that the annotation has been assigned due to evidence that an ancestral gene is involved in the biological process in question (i.e. linked to orthology) [36, 37]. None of the annotations with the IBA evidence group are annotations of *A. thaliana* proteins. In fact, only 13 annotations of proteins of other species than *A. thaliana* are annotated with an evidence group other than IBA. From this, it can be concluded that almost all cell death-related GO BP annotations, annotated by experts, and with evidence other than orthology, are annotations to proteins of a single species, *A. thaliana*.

Interpreting the state of cell death-related annotations to *Viridiplantae* proteins is highly dependent on what annotation evidence is accepted by the interpreter. This is mentioned because there is a large difference in the number of total annotations, and the number of species with annotated proteins, depending on whether the annotations have been performed by automatic or manual means. Moreover, the evidence used by experts to manually annotate proteins of plant species other than *A. thaliana* seems to almost exclusively be orthology evidence.




| GO term | GO term name |
|---|---|
| GO:0009626 | plant-type hypersensitive response |
| GO:0012501 | programmed cell death |
| GO:0010343 | singlet oxygen-mediated programmed cell death |
| GO:0008637 | apoptotic mitochondrial changes |
| GO:0006915 | apoptotic process |
| GO:0070059 | intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress |
| GO:0006309 | apoptotic DNA fragmentation |



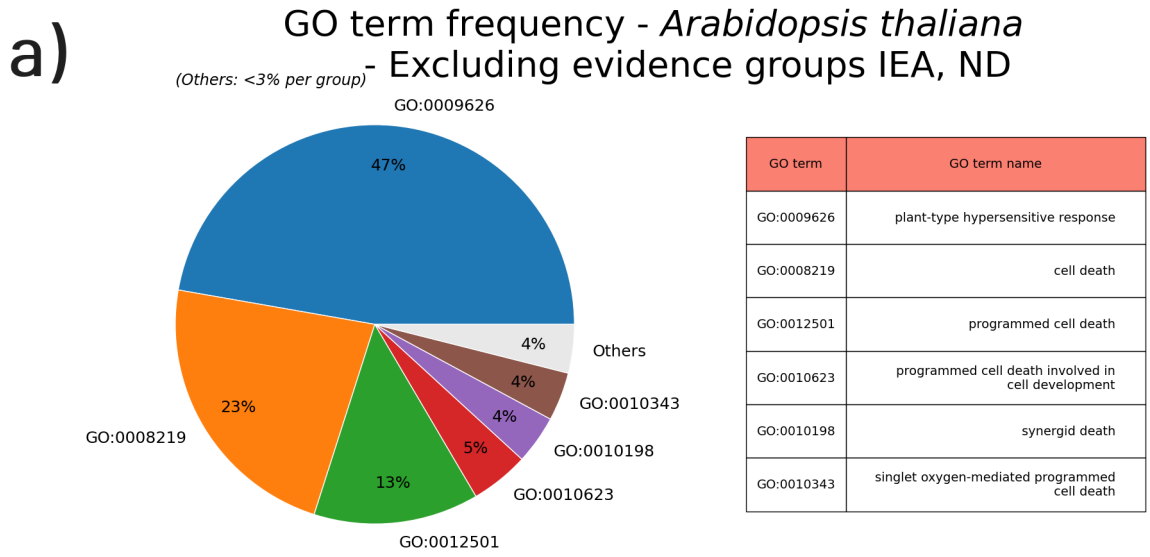

| GO term | GO term name |
|---|---|
| GO:0009626 | plant-type hypersensitive response |
| GO:0008219 | cell death |
| GO:0012501 | programmed cell death |
| GO:0010623 | programmed cell death involved in cell development |
| GO:0010198 | synergid death |
| GO:0010343 | singlet oxygen-mediated programmed cell death |

*Figure 3: Frequency of cell death-related GO BP terms in annotations of proteins of species in the Viridiplantae clade. (a) Frequency of terms when the annotations were based on all evidence types. The "Others" group contains all terms that had a frequency lower than 3%. (b) Frequency of terms when the annotations were based on evidence from all evidence groups except IEA and ND. The "Others" group contains all terms that had a frequency lower than 3%.*

### Cell death-related GO BP annotation of proteins of *A. thaliana*

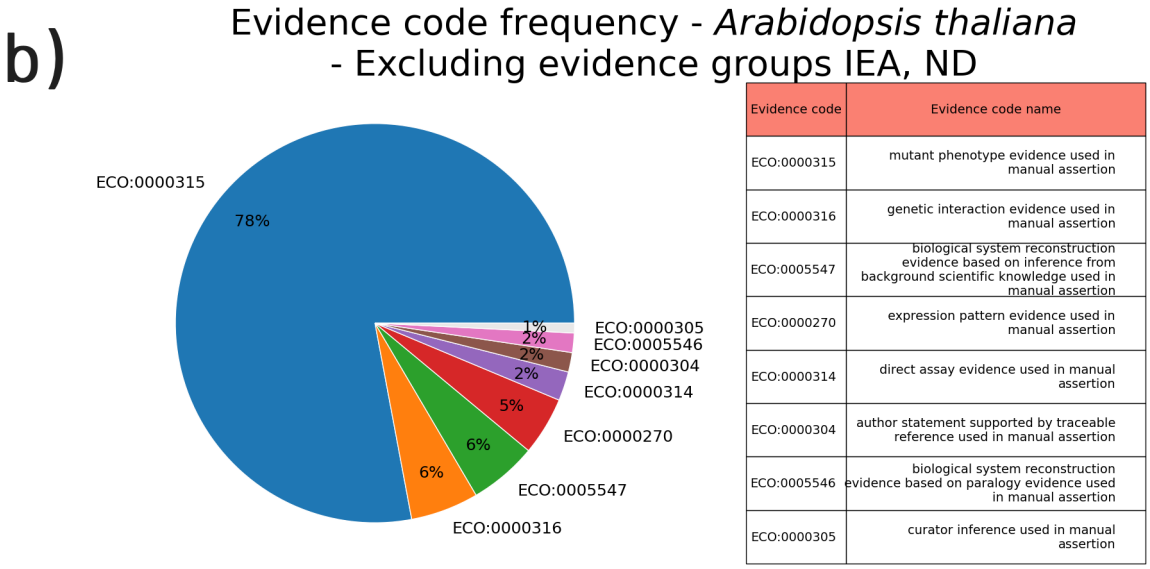
The number of cell death-related GO BP term annotations to *A. thaliana* proteins that have been manually annotated by experts is 127. These annotations had one of ten GO terms (full table (table 8) of terms and annotation frequency is given in the Appendix C.1). The GO term frequency can be seen in figure 4a. The most prevalent GO term is "plant-type hypersensitive response" (GO:0009626), followed by the higher-order GO terms "cell death" (GO:0008219) and "programmed cell death" (GO:0012501). Of the 127 annotations most (99 total) have been annotated with the evidence code of "mutant phenotype evidence used in manual assertion" (ECO:0000315), meaning that a phenotype (here likely to be cell death) was observed in units having a mutation of the protein in question. All the evidence codes used for these 127 annotations can be seen in figure 4b.

a)




| GO term | GO term name |
|---|---|
| GO:0009626 | plant-type hypersensitive response |
| GO:0008219 | cell death |
| GO:0012501 | programmed cell death |
| GO:0010623 | programmed cell death involved in cell development |
| GO:0010198 | synergid death |
| GO:0010343 | singlet oxygen-mediated programmed cell death |

b)



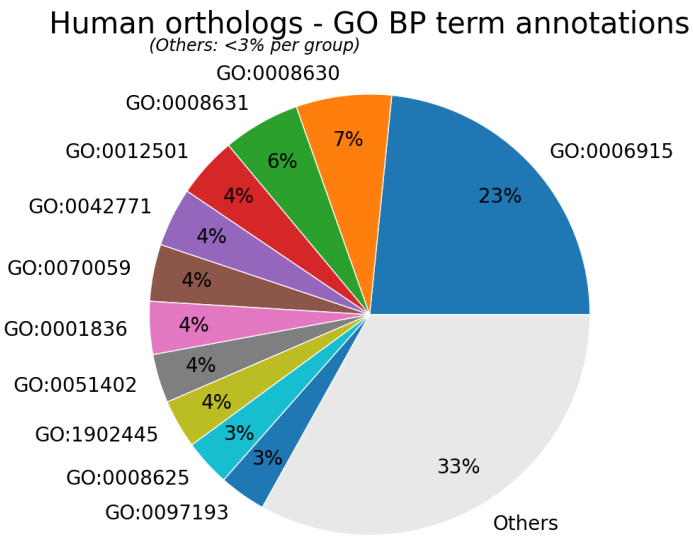

| Evidence code | Evidence code name |
|---|---|
| ECO:0000315 | mutant phenotype evidence used in manual assertion |
| ECO:0000316 | genetic interaction evidence used in manual assertion |
| ECO:0005547 | biological system reconstruction evidence based on inference from background scientific knowledge used in manual assertion |
| ECO:0000270 | expression pattern evidence used in manual assertion |
| ECO:0000314 | direct assay evidence used in manual assertion |
| ECO:0000304 | author statement supported by traceable reference used in manual assertion |
| ECO:0005546 | biological system reconstruction evidence based on paralogy evidence used in manual assertion |
| ECO:0000305 | curator inference used in manual assertion |

*Figure 4: (a) Frequency of cell death-related GO BP terms for annotations of proteins of A. thaliana. The annotations include all evidence groups except IEA and ND. The "Others" group contains all terms that had a frequency lower than 3%. (b) Frequency of evidence codes for in the annotations of cell death-related GO BP terms for annotations of proteins of A. thaliana. The annotations include all evidence groups except IEA and ND.*

## 7.2 *A. thaliana* orthologs of cell death-related GO BP term annotated proteins of other species

In total, 436 *A. thaliana* protein orthologs of cell death-related GO BP term annotated proteins of other species were identified using the methods of this project. Only three of these (RBOHD, RBOHE, and E2F3) had already been incorporated into the KN, with the two former having been manually curated and the latter incorporated due to its annotation (in *A. thaliana*) with a cell death-related GO BP term. This is an intriguing observation as I expected there to be more proteins that were described as involved in RCD in *A. thaliana* and had orthologs in other species where the orthologs were also involved in RCD in the given species.

Four hundred one of these 436 identified orthologs are orthologs of human proteins, while 35 are orthologs of proteins of *Viridiplantae* species. Of the 35 *Viridiplantae* orthologs, 31 are annotated with the GO term "programmed cell death involved in cell development" (GO:0010623) (with the evidence group ISS ("sequence similarity evidence used in manual assertion")), while 4 are annotated with the GO term "plant-type hypersensitive response" (GO:0009626) (with the evidence group IBA ("biological aspect of ancestor evidence used in manual assertion")). The human orthologs are annotated with many different cell death-related GO BP terms as can be seen in the figure 5. From the figure can also be observed that the most prevalent GO term annotations are terms describing processes of apoptosis. So, if the *A. thaliana* orthologs are connected to validated entities of RCD in *A. thaliana*, there may be functional similarities between entities involved in apoptosis in humans and entities involved in RCD in *A. thaliana*.




| GO term | GO term name |
|---|---|
| GO:0006915 | apoptotic process |
| GO:0008630 | intrinsic apoptotic signaling pathway in response to DNA damage |
| GO:0008631 | intrinsic apoptotic signaling pathway in response to oxidative stress |
| GO:0012501 | programmed cell death |
| GO:0042771 | intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator |
| GO:0070059 | intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress |
| GO:0001836 | release of cytochrome c from mitochondria |
| GO:0051402 | neuron apoptotic process |
| GO:1902445 | regulation of mitochondrial membrane permeability involved in programmed necrotic cell death |
| GO:0008625 | extrinsic apoptotic signaling pathway via death domain receptors |
| GO:0097193 | intrinsic apoptotic signaling pathway |

*Figure 5: Frequency of cell death-related GO BP term annotations of the human proteins of the A. thaliana orthologs in the knowledge network.*

## 7.3 The knowledge network

The knowledge network (KN) (figure 6) consists of 2 026 nodes and 9 407 links. Most nodes represent genes or gene products (i.e. proteins) (2 005 nodes), while a few represent other chemical substances (21 nodes). Of all the nodes, about 8% (157 nodes) represent manually curated entities, 4% (85 nodes) are *A. thaliana* proteins annotated with a cell death-related GO BP term, 21% (432 nodes) *A. thaliana* orthologs of proteins annotated with cell death-related GO BP term in another species, and 66% (1 328 nodes) proteins interacting with entities of the three other mentioned groups. The remaining 1% (24 nodes) represents entities that are incorporated

in the KN due to sharing co-expression or co-occurrence relations with the three first-mentioned groups. The links represent different relations, with 56% (5 266 links) co-expression, 22% (2 027 links) physical interactions, 15% (1 438 links) co-occurrence in abstracts, 7% (638 links) predicted interactions. The remaining links (<1%) represent manually curated interactions (30 links) and genetic interactions (8 links). Examples of annotation data accessible to the user when the KN is displayed in the yEd network software are presented in figure 7.

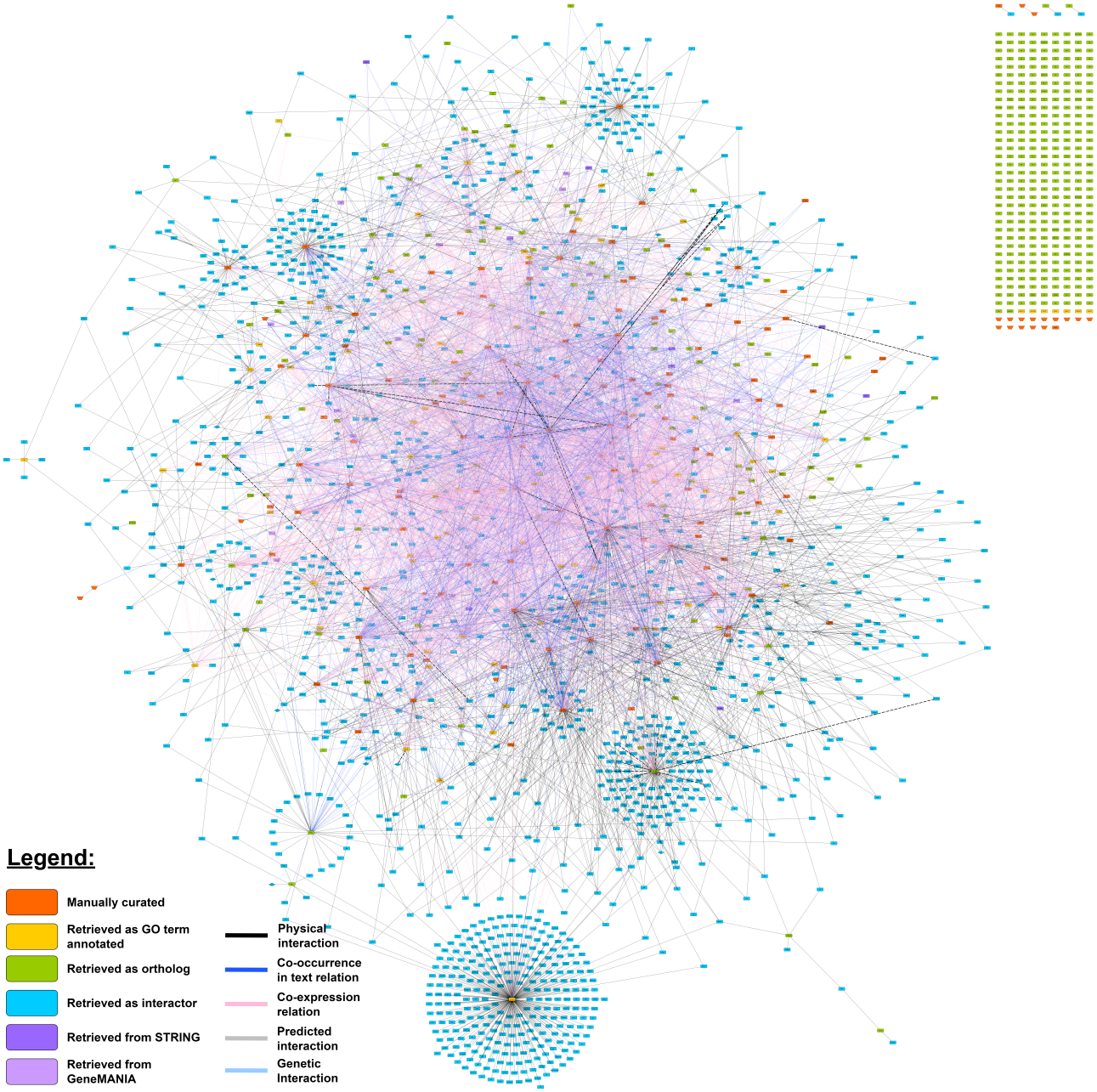



*Figure 6: Overview of the entire knowledge network. The node colors indicate on what basis the entity was incorporated into the network. The link colors indicate what type of relation the links represent. Entities that were manually curated or included in the network due to being annotated with cell death-related GO Biological Process terms can be considered validated entities of RCD in A. thaliana. In contrast, putative RCD entities are entities that were included based on orthology (i.e. A. thaliana orthologs of proteins annotated with cell death-related GO Biological Process terms in another species) or included as interactors with any of the other entities of the network (i.e. having an experimentally validated physical interaction). Entities retrieved from the databases STRING and GeneMANIA must also be considered as putative entities of RCD. The network is displayed with a layout that puts nodes that are more interconnected in the center. With this layout, the validated entities of RCD tend to reside more toward the center of the network than other entities. In the upper right corner are shown entities that have no relations connecting them to the main network. Eight of these are however connected in pairs. Dashed links signify physical interactions that are either phosphorylation or dephosphorylation reactions.*

### Unconnected nodes

Most of the nodes are connected in what will be referred to as the main network, while a smaller number (338 nodes) are not. In the latter group, eight nodes are connected in pairs, and 330 nodes are not connected by any

| Data | |
|---|---|
| URL | https://www.uniprot.org/uniprotkb/P94077/entry |
| Description | |
| id | ent:104 |
| name | Protein LSD1 |
| abbreviation | LSD1 |
| synonyms | |
| molecule_type | protein |
| species | Arabidopsis thaliana |
| UniProt_accession | P94077 |
| TAIR | AT4G20380 |
| PubChem | |
| InterPro | IPR040319\|IPR005735 |
| OrthoDB | NA |
| GO_term_Cell_Death | |
| GO_term_other_BF_annotations | |
| cellular_compartments | cytoplasm\|nucleus |
| found | Manually curated |
| post_translational_modifications | |
| URL | https://www.uniprot.org/uniprotkb/P94077/entry |
| note | |
| reference | PMID:36179088\|PMID:29495279 |
| curator_note | Negatively regulates RCD propagation (PMID:361... |
| article_type | review |
| InterPro_family | IPR040319(LSD1-like) |
| InterPro_homologous_superfamily | |
| InterPro_ptm | |

*(a)*

| Data | |
|---|---|
| URL | |
| Description | |
| Confidence values | 0.56 |
| Special interaction | phosphorylation reaction |
| IntAct accession | EBI-8575169\|EBI-8574940 |
| Taxid interactor A | Arabidopsis thaliana |
| Taxid interactor B | Arabidopsis thaliana |
| Taxid interactor A code | 3702 |
| Taxid interactor B code | 3702 |
| Interaction types | direct interaction(EBI-8574940)\|phosphorylation r... |
| Interaction types code | MI:0217(EBI-8575169)\|MI:0407(EBI-8574940) |
| Detection methods | pull down(EBI-8574940)\|protein kinase assay(EBI... |
| Detection methods code | MI:0096(EBI-8574940)\|MI:0424(EBI-8575169) |
| Host organisms | In vitro(EBI-8575169;EBI-8574940) |
| Host organisms code | -1(EBI-8574940;EBI-8575169) |
| References | pubmed:21276203(EBI-8574940;EBI-8575169) |
| Source databases | MINT, Dpt of Biology, University of Rome Tor Verga... |
| found | IntAct |
| interaction | Experimentally validated |
| curator_note | |

*(b)*

*Figure 7: Screenshot of the annotation data as seen when displaying the knowledge network in the yEd network software. (a) Example of node annotation data. (b) Example of link annotation data. Empty cells indicate that the annotation was not retrieved, not relevant, or not available. If the annotation was not available from the source database the annotation could also be stored as "NA". Multiple annotations of the same category were separated by "\|".*

links. Of the unconnected nodes, some are relevant to RCD as they were either manually curated (e.g. PAMPs, hormones, ions, and non-biologically derived inhibitors of RCD) or because they were annotated with cell death-related GO BP terms. These are likely to actually be connected within the network, but the connecting relations were (for some reason) not acquired in this project, which they should be for future projects. These entities are therefore retained as a reminder for future projects. For the unconnected protein nodes, 313 out of 316 have accession codes linking to the TrEMBL part of the UniProt database. This means that almost all are characterized as unreviewed with limited data. Moreover, most of these are *A. thaliana* orthologs of proteins annotated with a cell death GO BP term in another species. The orthologs in this group are thereby putative entities of plant RCD based solely on orthology, making their putative role less likely than orthologs in the main network that also have some sort of relation to other entities.

In the unconnected group, there are isoforms of proteins in the main network, i.e. they have the same gene locus code but different UniProt accession codes. These unconnected proteins are important because they have annotations stating their role (with evidential value) in plant RCD, which the isoforms do not have. For example, the LSD1 gene (gene locus code; AT4G20380) has two isoforms in the KN, with one isoform connected in the main network (UniProt:P94077) and the other not (UniProt:F4JUW0). The former is included in the KN because it is an *A. thaliana* ortholog of cell death-related GO BP term annotated protein in another species, while the latter is included due to being a *A. thaliana* protein annotated with a cell death-related GO BP term. It could be that the GO annotation and orthological link are suited for both isoforms (e.g. if the annotation was meant for the gene level), which can only be determined with information on the annotation reasoning, which was not acquired in this project. In this example, removing F4JUW0 would mean removing the information that on the basis of GO annotation LSD1 is described as part of plant RCD. The same example applies to BAK1 (AT4G33430).

### 7.3.1 Network analysis

Network analysis was performed to get the network and node properties. These properties may indicate certain interesting aspects of the network. The following section will present the network analysis of the entire KN with all the various relation types found between the entities. In the analysis, the KN is treated as an undirected and unweighted network. This means that all types of relation links are considered to have equal influence on the network properties. The reader must also be aware that relations other than physical interactions were not

retrieved for the proteins incorporated into the KN based on being interactors with the other proteins of the KN. The rationale behind this decision will be discussed in another section. Consequently, the choices made for the methods used to generate the KN impact its properties, like the node degrees.

**Node degree and degree distributions**

Network properties are often presented in the context of the node degrees, i.e. how many links are connected to any given node. In the KN, most nodes have a low degree, with 68% (1 374 out of 2 026 nodes) having a degree of 0 or 1. Only a few nodes have a particularly high degree. There are for instance only 29 entities with a degree above 100. NAC089, MPK3, SOBIR1, BAK1, and BIR1 have the highest node degrees and are thereby the largest hubs of the KN. The distribution of the node degrees follows a power-law distribution as there is a strong linear relationship between the logarithmically transformed node degree values and the logarithmically transformed frequency values. The linear relationship was calculated using Ordinary Least Squares (OLS) regression, which yielded an $R^2$ value of 0.7. A network with a power-law degree distribution is characterized as a scale-free network, which is typical for biological networks. The degree distribution for the KN is shown in figure 8.

Grouping the nodes of the KN according to how the nodes were retrieved for the KN can show interesting aspects of the individual groups. For instance, figure 9 shows that the degree distribution is not the same for the individual groups. Entities manually curated and proteins annotated with cell death-related GO BP terms tend to have higher degrees (i.e. be more connected) than entities of the other groups. The median degree value for these respective groups is 52 and 43. The nodes of these groups are considered particularly important as players of RCD in *A. thaliana*. As their involvement in RCD is well described it is reason to suspect that the high degree is influenced by many co-occurrence relations between other well-described RCD players. An observer must therefore be aware that a hub in the KN may not serve as a functional hub *in vivo*. The groups of nodes considered to be more putative entities of RCD, i.e. the group of proteins that were included based on orthology and the group of proteins that had physical interactions with other entities of the KN, had a lower median degree than the two former. They respectively have a median degree of 28 and 1. To conclude, the entities that are already deemed to be part of RCD tend to be more connected and form hubs in the KN than putative RCD entities.

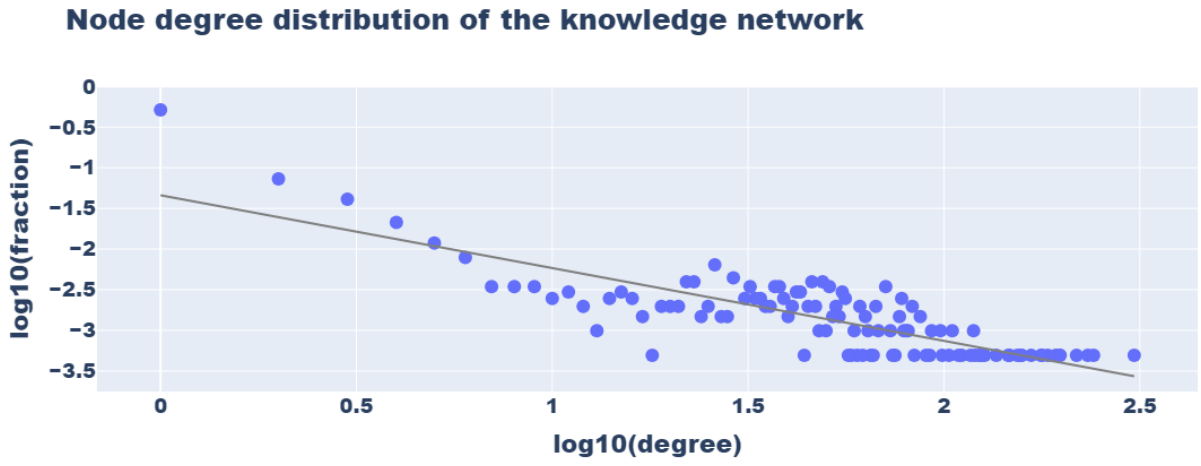



*Figure 8: The distribution for how connected the nodes are in the knowledge network, i.e. the degree distribution. The fraction of nodes with a given connectivity (i.e. degree) is shown over the same connectivity value. Nodes in the knowledge network with degree 0 were excluded from this graph. Both variables have been logarithmically transformed with base 10. The trendline is the Ordinary Least Squares regression line with $R^2$ of $\sim$0.7.*

The retrieval of all relation types for all entities could have shifted the overall degree distribution for the KN. However, as a PPI network is expected to be scale-free, the incorporation of data retrieved from IntAct (which

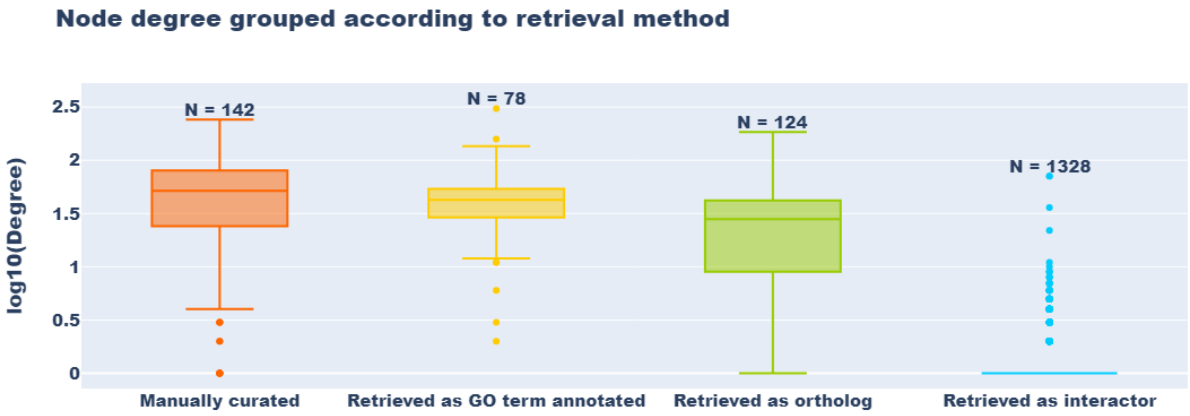
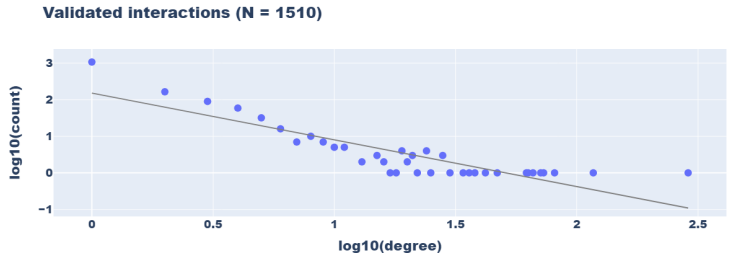



*Figure 9: Box plots indicating the spread of degree-values for nodes (i.e. number of connections to a node) grouped according to how they were retrieved and incorporated in the knowledge network (figure 6). Nodes with a degree less than 0 were removed from the groups. The degree values have been logarithmically transformed with base 10. (Red) Manually curated entities retrieved from the literature on plant RCD. The manually curated proteins could also be annotated with cell death-related GO Biological Process terms. (Yellow) Proteins retrieved by programmatic means due to being annotated with cell death-related GO Biological Process term(s). Proteins were identified in the QuickGO database. (Green) Retrieved as putative proteins in RCD of A. thaliana based on cell death-related GO Biological Process term annotation of orthologous genes in other species. Proteins were identified in the OrthoDB database. (Blue) Retrieved as putative proteins in RCD of A. thaliana due to having an experimentally validated physical interaction with proteins of one of the other mentioned groups. Proteins were identified in the IntAct database. The logarithmic median values for the respective groups are 1.7, 1.6, 1.4, and 0. The number of entities within each group is indicated (N). The figure does not show the groups of comparatively few nodes retrieved from the GeneMANIA and STRING databases (respectively 15 and 9 nodes).*

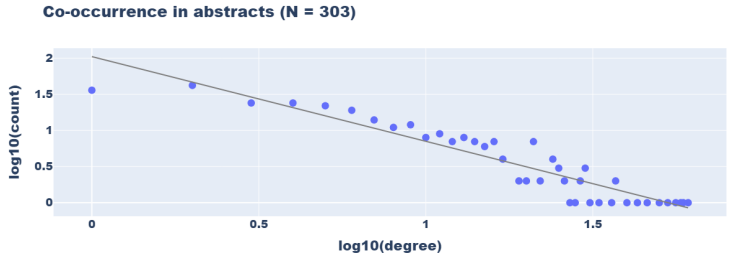
effectively forms a PPI network) should not remove the scale-free property of the KN. As can be seen from figure 10 the subsections of the KN where entities are solely connected by a specific relation type also have the scale-free property (due to the linear relationship between the variables), except from the subsection with co-expression relations. It can therefore be expected that the retrieval of other relation types to the proteins retrieved due to having a physical interaction with other proteins of the KN would result in connections within this group and to other entities of the KN also in a scale-free manner.

As mentioned in figure 10 the subsection of the KN consisting of nodes connected by co-expression the degree distribution does not follow a power-law distribution. This subsection should therefore not be categorized as scale-free. With closer inspection of the KN was observed that certain pairs of nodes were connected to each other by more than one co-expression link. This can be a flaw in the data pipeline which might have affected the mentioned degree distribution in this subsection.

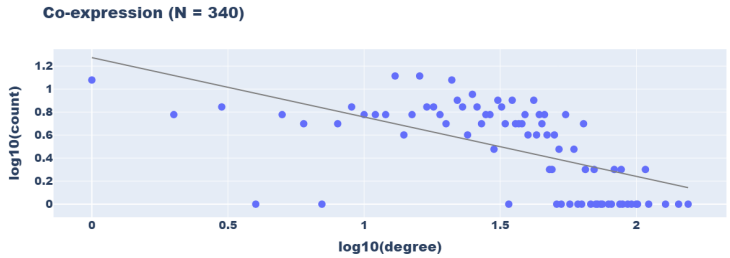
Some interesting aspects can be drawn from the group of nodes consisting of proteins included in the KN due to having physical interaction(s) with other entities of the KN (i.e. those retrieved from IntAct). In this group, 77% (1 028 out of 1 328) had a degree of one. The prevalent low degree is undoubtedly caused by the mentioned choice of not retrieving other relations for this group, as there are likely other relations that would have increased the degree of certain nodes. A degree of 1 reveals that most of the proteins in this group do not have interactions with more than one entity of the network. Thereby, most do not form hubs. However, those interactor proteins with higher degrees can be very important as hubs. For instance, NHL3 (UniProt:Q9FNH6) and CNIH1 (UniProt:Q9C7D7) have the degrees 11 and 10, and can then be considered to form physical interaction hubs. These interaction hubs can play an important role in RCD, and should therefore be experimentally assessed in the context of RCD. A table of nodes within this group having a degree greater than five is given in Appendix C.2 (table 9).



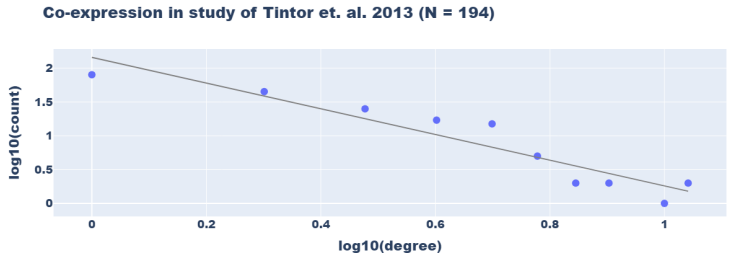

*(a)*




*(b)*




*(c)*




*(d)*

*Figure 10: Degree distribution for subsections of the knowledge network, where the subsection is extracted by only retrieving the specified relation (in figure title) and the connected entities. The number of nodes in subsections (N) is indicated in the figure title. (a) $R^2$ is $\sim 0.77$. (a) $R^2$ is $\sim 0.88$. (c) $R^2$ is $\sim 0.36$. (d) $R^2$ is $\sim 0.89$. The subfigure (d) is the degree distribution of entities that had co-expression relations identified in an immunity study that was deemed to have the experimental conditions closest to RCD [38].*

**Average clustering coefficients**

Groups of nodes that form clusters due to their interconnectivity are of general interest. However, the types of relations interlinking a group matter for the interpretation. For instance, a group being fully interconnected (i.e. forming a clique) by physical interaction links may be derived from a protein complex. It is important to remember that although groups of nodes form cliques in the static KN, the relations may only occur in certain scenarios, i.e. the temporal and spatial aspects always need to be accounted for in the context of biological systems. For instance, proteins that in theory can form a complex may not do so *in vivo* as the individual proteins may not be present together at any given moment, e.g. through different cellular localization or due to opposing expression levels. Similarly, co-expression of a group of genes may not occur if the combinatorial regulation for all genes facilitates it. Even though this must be accounted for, clusters of highly interconnected nodes are key areas of the KN that may be particularly interesting when acquiring knowledge about the system. The clustering coefficient of a node is one of the metrics that can imply the presence of clusters.

Regarding the clustering coefficients (CC) for the nodes of the KN, there is a trend that nodes with a relatively low degree have a relatively high CC, whereas nodes with a high degree have a low CC. Figure 11 shows the average CC for every degree value. The distribution of the average CC indicates that there is only a weak linear relationship between the average CC and the degree (degree as logarithmically transformed). The OLS regression line was calculated to have an $R^2$ value of only 0.16. The low linear relationship indicates that the KN can not be considered a true hierarchical network type.

The nodes with a degree of five had the highest average CC of 0.55. Within this group of nodes, eight had the maximum CC value of 1.00. These eight nodes are therefore all part of cliques. Moreover, they were all retrieved for the KN as interactors of other entities in the network (i.e. retrieved from IntAct). It turned out that these eight nodes formed individual 6-cliques (i.e. cliques with six nodes) with a group of five nodes that themselves formed a 5-clique. In other words, all of the eight nodes were connected to all nodes in the 5-clique, but none of the eight nodes were connected to each other. Interestingly, the connections the eight nodes had to the 5-clique were all physical interactions. The 5-clique itself is interesting as all proteins are deemed to be part of RCD (i.e. four were manually curated, while one was retrieved for the KN due to being annotated with a cell death-related GO BP term). The 5-clique forms a clique on both the basis of co-expression and co-occurrence relations, but not physical interactions (or other relations). A sketch to graphically show what has here been described is given as figure 12. More information regarding the nodes forming the mentioned 5-clique is given in Appendix C.3 (table 10). In Appendix C.3 is also given similar information on the eight mentioned nodes (table 11). The reason this discovery is interesting is because; all of the eight proteins interact with all members of the group of five nodes, which itself is so interconnected and relevant in RCD.

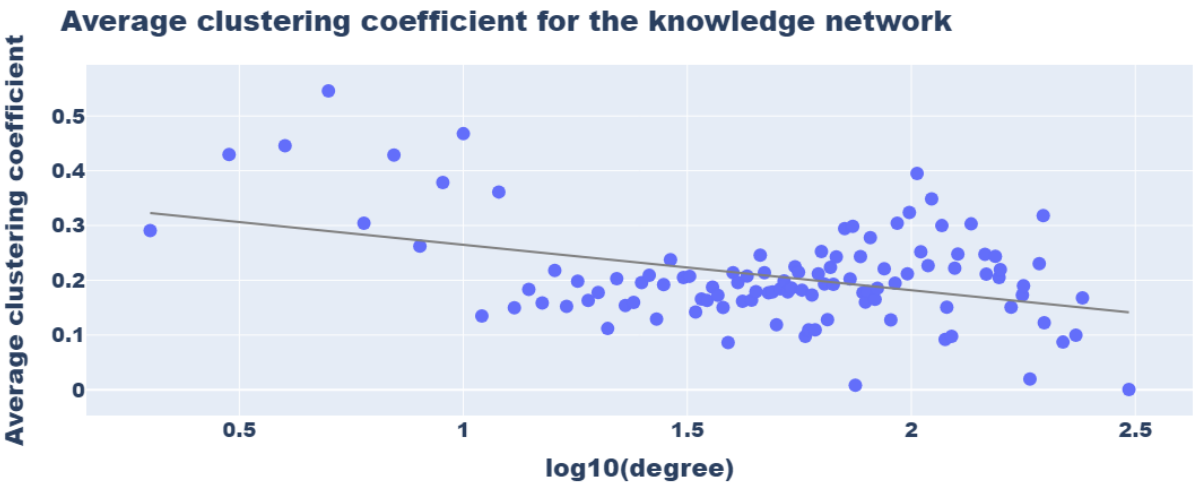



*Figure 11: Average clustering coefficient (CC) over logarithmically transformed degree. A high CC signifies that a given node's neighborhood is highly interconnected. The average CC value is the mean CC value for all nodes with the same degree (i.e. the same number of connections). The trendline is the Ordinary Least Squares regression line with an $R^2$ of ∼0.16, thereby indicating a weak linear relationship between the average CC and the logarithmically transformed degree value.*

## 7.4 Use cases

### 7.4.1 Using the KN and plant RCD studies in combination

The KN is a collection of information from multiple sources, thereby serving as a prior knowledge bank. The prior knowledge is important for the interpretation of new research findings. For instance, a study can present transcriptomic data that by itself brings some value to the discussion. However, if the transcriptomic data is put into context with what is already known, like interaction events, this can give more information about the consequences of the gene expression levels to the system. With the varying data types in the KN, it can serve as a great first approach for researchers when they want to interpret the results of studies in a broader context. Moreover, it can help to prioritize which entities to interpret first. For instance, all entities that are highlighted in an experimental study and are represented in the KN have a likelihood of being involved in plant RCD, at least a higher likelihood than those entities that are not found in the KN. It does not mean that one should discard all highlighted entities from the research findings, but rather that these should be ranked second in the interpretation queue. The following use case is meant to present these thoughts.

Burke et. al. conducted a study looking at transcriptomics during RCD induced by salicylic acid (SA), heat shock (HS), and critical culture dilution [39]. They identified 11 genes that were differently expressed genes (DEGs) in all tested conditions. One of these 11 genes is present in the KN presented in this project. The gene is HSR4 (AT3G50930), which was included in my KN because it is annotated with cell death-related GO BP terms ("cell death" (GO:0008219) and "plant-type hypersensitive response" (GO:0009626)). In this study, they concluded that only a few genes are DEGs under all tested conditions, leading to the hypothesis that potentially few genes are regulated similarly across different plant RCD processes. For the SA-induced condition, 83 out of 1 173 DEGs found in the study were also found in the KN, 3 out of 59 for the HS condition, and 3 out of 16 for the critical dilution. To show how the KN can be useful for further experiments I will discuss the 3 DEGs in the HS-induced RCD condition.

Figure 13 shows a subsection of the KN, consisting of the three genes, ATHB-9 (AT1G30490), GRF2 (AT1G78300), and VPEG (AT4G32940), that are the same genes identified as upregulated during the HS condition in the mentioned study. In the figure, the three genes are also shown with their respective neighborhoods in the KN. Both ATHB-9 and GRF2 were included in the KN because they share physical interactions with some other entity of the KN. As a consequence, they are only considered putative RCD entities with fairly limited evidence

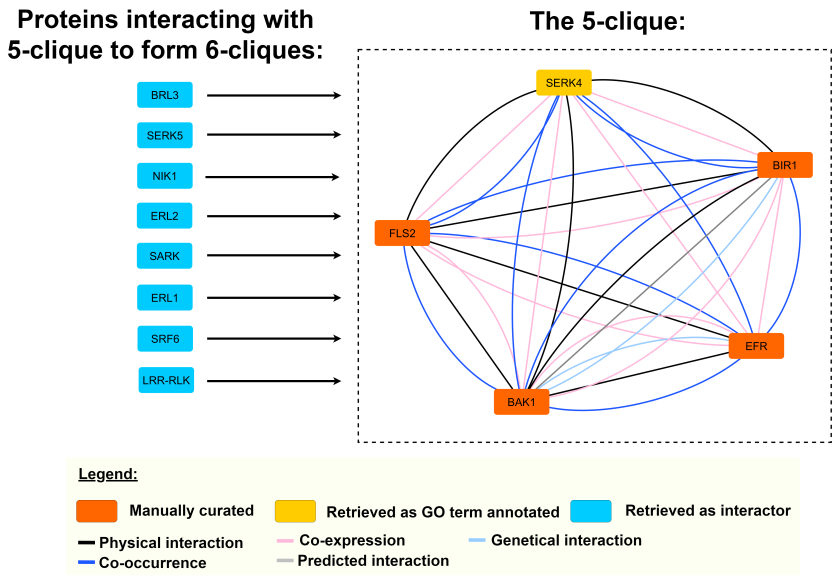



*Figure 12: Sketch to show an interesting set of proteins that were deemed most interesting in the knowledge network because of their interconnectivity. The sketch shows the nodes (blue nodes) that are individually connected with every member of the group of 5 proteins (nodes within the dashed area) that by themselves are fully interconnected. A fully interconnected group of nodes is known as a clique. The blue nodes are individually connected to the 5-clique so that 6-cliques are formed. In this sketch, the singular arrows are meant to represent all the physical interaction connections the individual blue nodes have to the 5-clique to form a 6-clique. The relation types within the 5-clique are shown to display that both co-occurrence and co-expression relations on their own give rise to the clique. For this example, multiple copies of the same relation type between two nodes were excluded so that only one is shown. Physical interaction self-loops were also excluded.*

(i.e. only based on the link with physical interaction). In the KN, the interactions that ATHB-9 and GRF2 share with other proteins in the KN have been described with low evidence strength, as will be discussed further. ATHB-9 is connected by physical association interactions (MI:0915), meaning that the proteins are in the same physical complex but may not be in direct contact. GRF2 has interactions characterized with even less specificity, being association interactions (MI:0914).

In UniProt, ATHB-9 is described as a probable transcription factor (TF). ATHB-9 is shown to interact with the other TFs AGL63 (AT1G31140) and HEC3 (AT5G09750). These interactions were identified in studies looking at other biological processes than RCD [40, 41]. Here we get an example of how knowledge derived from some unrelated process may shed light on something that may also occur in the process we are interested in. However, as the knowledge is derived from another biological process experimentation is needed to identify if the interaction also occurs as part of RCD. Both interactions were identified solely using a yeast two-hybrid system, meaning that experimental methods are needed to determine if the interaction also occurs *in planta*. However, the evidence of upregulation provided by the Burke study does by itself strengthen ATHB-9's position in the KN as a putative RCD player. Based on all this information I would propose performing experiments on the role of ATHB-9 in RCD (especially in HS conditions). These experiments should reveal if ATHB-9 affects RCD, and if so, are the interactions between ATHB-9 and AGL63 or HEC3 the cause for the effect on RCD. Co-immunoprecipitation could be used to identify if the proteins form a complex during induced RCD conditions, but it will not reveal if the complex formation itself affects RCD. Performing a gene knock-out (KO) of ATHB-9 and comparing the phenotype with a control when RCD is induced could reveal if ATHB-9 is involved in RCD. Other experiments should also be performed to increase the evidence.

The transcription regulator GRF2 was in a study found to interact with, among others, five of the proteins found in the KN. These were CAT2 (AT4G35090), ACS6 (AT4G11280), CPK1 (AT5G04870), BAK1 (AT4G33430), and GRF6 (AT5G10450) [42]. These interactions were identified using tandem affinity purification (MI:0676), with GRF2 as the bait, which gives little information other than that GRF2 forms a complex with these proteins *in vivo*. Again, determining their role in RCD processes can be done through gene-KO studies. The

interactions are however quite interesting in that they connect so many entities that were manually curated (all but GRF6) for the KN. As the manually curated proteins are expected to influence RCD an interaction with these may also be relevant for RCD. The interaction with GRF6, another transcription regulator, is interesting, also because GRF6 is co-expressed with VPEG (in a study on seed composition [43]) which also was upregulated in HS-induced RCD.

VPEG was manually curated for the KN, but has no physical interactions with other entities, although it is fairly connected by other relations. VEPG's role as a protease and its quite prominent occurrence in the RCD literature makes it a target for experimental studies for the identification of its physical interactions [44, 45, 46]. Moreover, it is co-expressed with many other proteins in the KN, although they were not identified as DEGs in the HS-induced RCD experiment of Burke et. al.

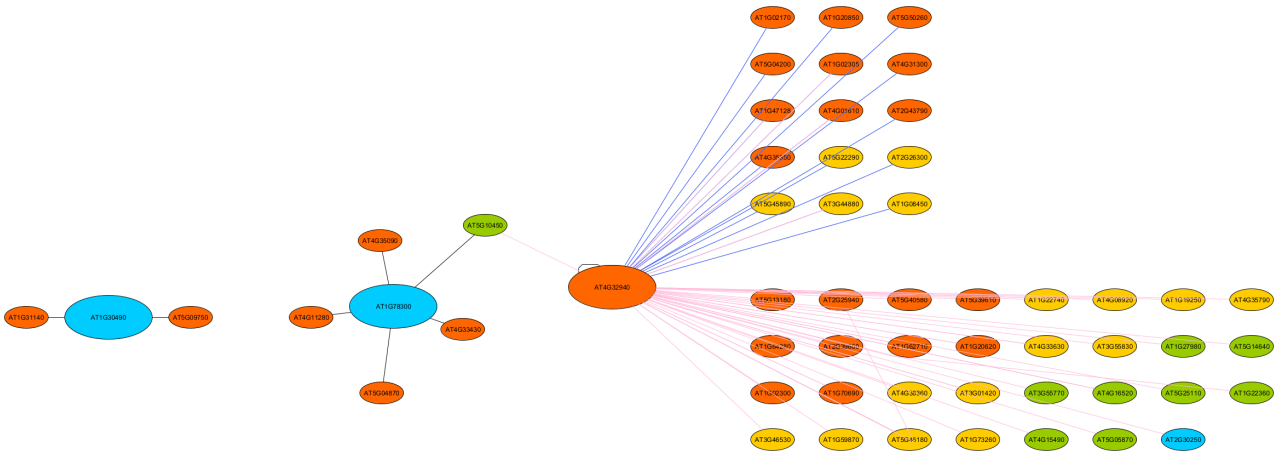


*Figure 13: Subsection of knowledge network showing the three entities (big nodes) that are also found as DEGs in the study of Burke et. al. [39] in an HS experiment. The three common entities are shown with the entities they connect with in the KN. In the HS experiment of the study, the three genes are upregulated.*

### 7.4.2 Knowledge retrieval and hypothesis generation by the use of the knowledge network

**Co-occurrence in abstracts**

Entities mentioned in the same abstract are likely to have a biological connection. Manually identifying these co-occurrences is time-consuming and unnecessary as there are automatic methods of doing this. STRING identifies co-occurrences in PubMed abstracts through text mining methods. A co-occurrence can provide valuable information, although there are limitations to the usefulness that one needs to be aware of. Despite the limitations, the co-occurrence relations provide a targeted approach to acquiring information. The following will serve as an example of how one could use the KN and the co-occurrence relations to identify proteins that should be assessed further to identify if they have descriptions of being involved in RCD processes. If there are descriptions of how they are involved, the proteins should be suggested as targets of annotations with appropriate RCD GO BP terms. Until the time they are annotated, a person generating newer versions of the KN may choose to manually curate these proteins as proteins involved in RCD.

From the KN is extracted a network that only contains co-occurrence relations and the connected proteins. Moreover, the network will only show co-occurrence relations between validated RCD proteins and proteins that were included in the KN based on orthology. In this example, proteins that are considered validated RCD proteins are those that had been manually curated and those that had been included in the KN due to being annotated with a cell death-related GO term. The example will not show co-occurrence relations found in-between proteins of the two groups, as they do not provide information on the point that this example tries to convey. The network is further organized to highlight what type of validated RCD protein every putative RCD protein has a co-occurrence with. The result is the network presented in figure 14. In this example, it is seen that some putative RCD proteins have co-occurrence(s) with manually curated proteins (i.e. the lower left group),

others have co-occurrence(s) with GO term annotated proteins (i.e. lower right group), and some that have co-occurrence(s) with both of these (i.e. the lower middle group). This latter group is of particular interest as they have co-occurrence relations to both manually curated proteins and proteins annotated with cell death-related GO BP terms. This group contains proteins with features likely relevant to RCD, like calcium-binding (CML50) (calcium flux in RCD), calcium transporters (ECA1, ECA2, ECA3, ECA4), respiratory burst oxidases (RBOHB and RBOHJ) (i.e. potentially involved in ROS production in RCD), superoxide dismutases (MSD1 and FSD2) (ROS degradation), a phosphatase (PP2AA3) (dephosphorylation), mitochondrial GTPase (MIRO3) (molecular switch), transcription factors (GRF6), a histone acetyltransferase (HAC12) (gene expression regulation), and one of the entities with the highest degree in the KN (KIN10). The lower left and right groups in the figure are interesting for the same reason, although they have co-occurrences with either manually curated entities or GO term annotated proteins and not both. The reader must be aware that although other types of relations and co-occurrences between proteins within the groups are not shown in the network of this example, they may be present in the original KN. To conclude, based on co-occurrence the orthologs shown in this example should be assessed further to inspect if there are descriptions in the literature of their involvement in RCD. Moreover, the user is advised to use the KN to inspect how the putative RCD entities of this example are otherwise connected in the KN.

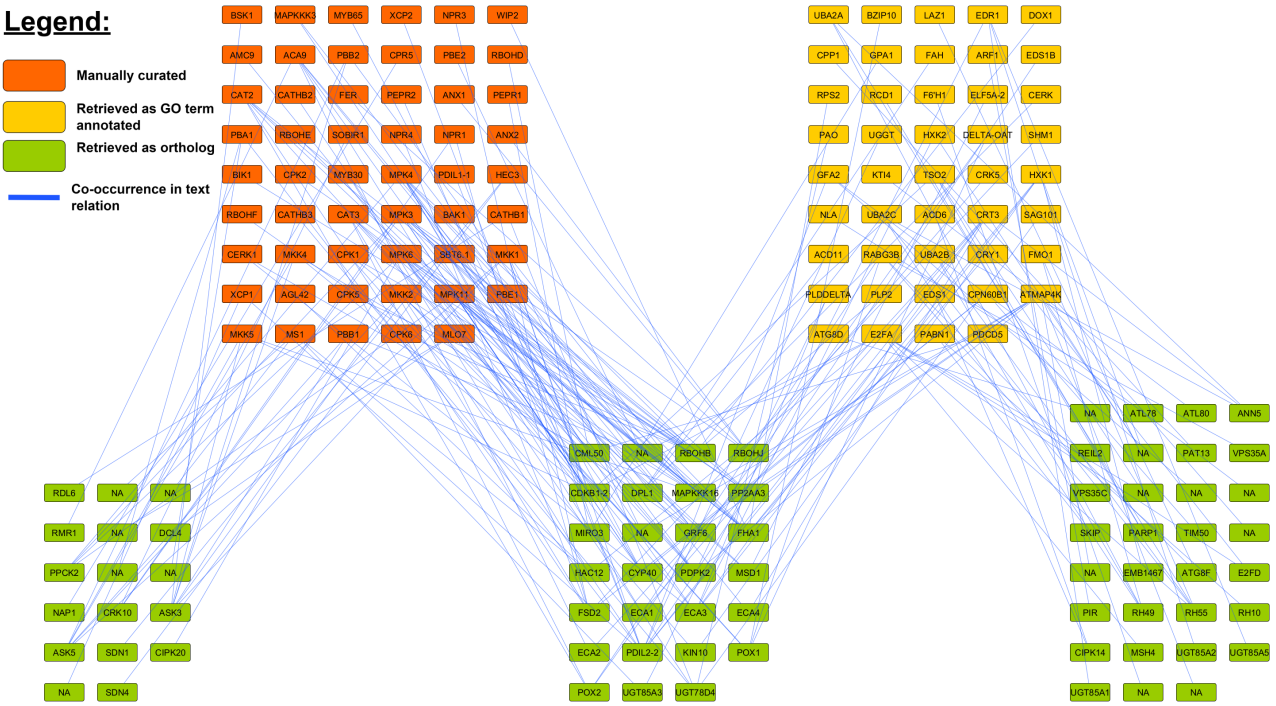



*Figure 14: A network derived from the knowledge network and shows putative RCD entities that are connected to validated RCD entities by co-occurrence relations in PubMed abstracts. The putative entities (green nodes) are A. thaliana orthologs of proteins annotated with cell death-related GO Biological Process terms in other species. The validated nodes are manually curated entities (red nodes) or entities annotated with cell death-related GO Biological Process terms (yellow nodes). In the derived network the co-occurrences within entities of the mentioned groups have been removed to minimize clutter. The label "NA" (i.e. not available) signifies that the abbreviated gene name was not found by the methods used in this project (i.e. the Python code). Other identifiers are present for these nodes in the knowledge network file.*

## 7.5 The data pipeline

The data pipeline has been organized to utilize the developed Python package, which incorporates methods of acquiring data from many bioinformatical databases. The data is acquired separately through the REST APIs of the individual databases. The raw data is transformed for greater human accessibility. In the pipeline, the data is transformed to conform to the format chosen for this project, which were tables for respectively node data and link data. The organization of the code, with sub-packages devoted to a specific database, means that the user has the flexibility to utilize only what is needed for their project. Moreover, with some alterations, the pipeline can be used with various inputs, e.g. the user's own curated data, other species, or other GO terms.

# 8 Discussion

The first objective of this project was to assess the current state of cell death-related GO BP term annotations to proteins of species in the taxonomic clade *Viridiplantae*, and more specifically proteins of *A. thaliana*. The assessment was conducted because the findings would directly influence what annotation evidence was accepted for annotated proteins to be included in the knowledge network (i.e. parts of the second objective). As automatic annotation methods had annotated *Viridiplantae* proteins with GO BP terms that describe processes not observed in plant RCD (i.e. apoptotic processes), the assessment culminated in the decision of only retrieving proteins manually annotated (i.e. annotated by experts).

The second and main objective of this project was to create a knowledge network (KN) of validated and putative entities of RCD in *A. thaliana*, and their relations. The project resulted in a KN that includes many entities that must be considered putative players in RCD of *A. thaliana*. The relations between the entities imply how the entities are connected, and thereby how they may be involved in the overall biological process of plant RCD. The KN is a resource to consult to prioritize new research findings, for instance, gene lists that are the result of an experiment. The KN also provides a resource to acquire information, that for instance can be used in the process of making polished process diagrams of plant RCD.

The third objective of presenting use cases for the KN was also achieved. The use cases presented here showcase that the KN can be used to retrieve information on entities and their relations that have been discovered but not necessarily interpreted in the context of RCD. It thereby highlights entities and relations that should be looked further into, regarding what more information may be available in the literature as well as what needs to be experimentally tested further.

The fourth objective of formulating a data pipeline to generate KNs was in many ways accomplished. As with any method, there is room for improvement. There are improvements to be made in the way the data is retrieved and processed to be part of the KN. For instance, there are annotation data of the entities that is retrieved from the databases using the current methods but some is processed in a manner that makes it available as part of the KN. Moreover, there are other biological databases that house even more information that should be retrieved for the KN. For future projects, the data pipeline should be improved as the way the KN is generated directly affects what it contains and what it can be used for. The Python code underlying the pipeline can be improved in areas to increase the readability of the code. Better readability and continuity in the code increase the likelihood of further development.

## 8.1 Discussion of results

Unfortunately, the proteins highlighted in the results as potentially interesting for the process of RCD have not, and will not be discussed further. As mentioned earlier, the interpretation of data requires a lot of domain-specific knowledge that I feel I do not currently possess. However, this also serves as an encouragement for further collaboration and the development of improved methods.

### 8.1.1 Formulating a decision tree to prioritize new research findings

One of the use cases shown for the KN was to prioritize experimental research findings according to how likely an entity of interest is to be involved in RCD. To perform this prioritization it might be beneficial with some guidance. As figure 15 is presented a decision tree

## 8.2 Limitations with the KN

There are limitations to the usability of the current version of the KN. Some of the limitations stem from the quality, quantity, and type of available data, while others stem from the methods used in the construction of

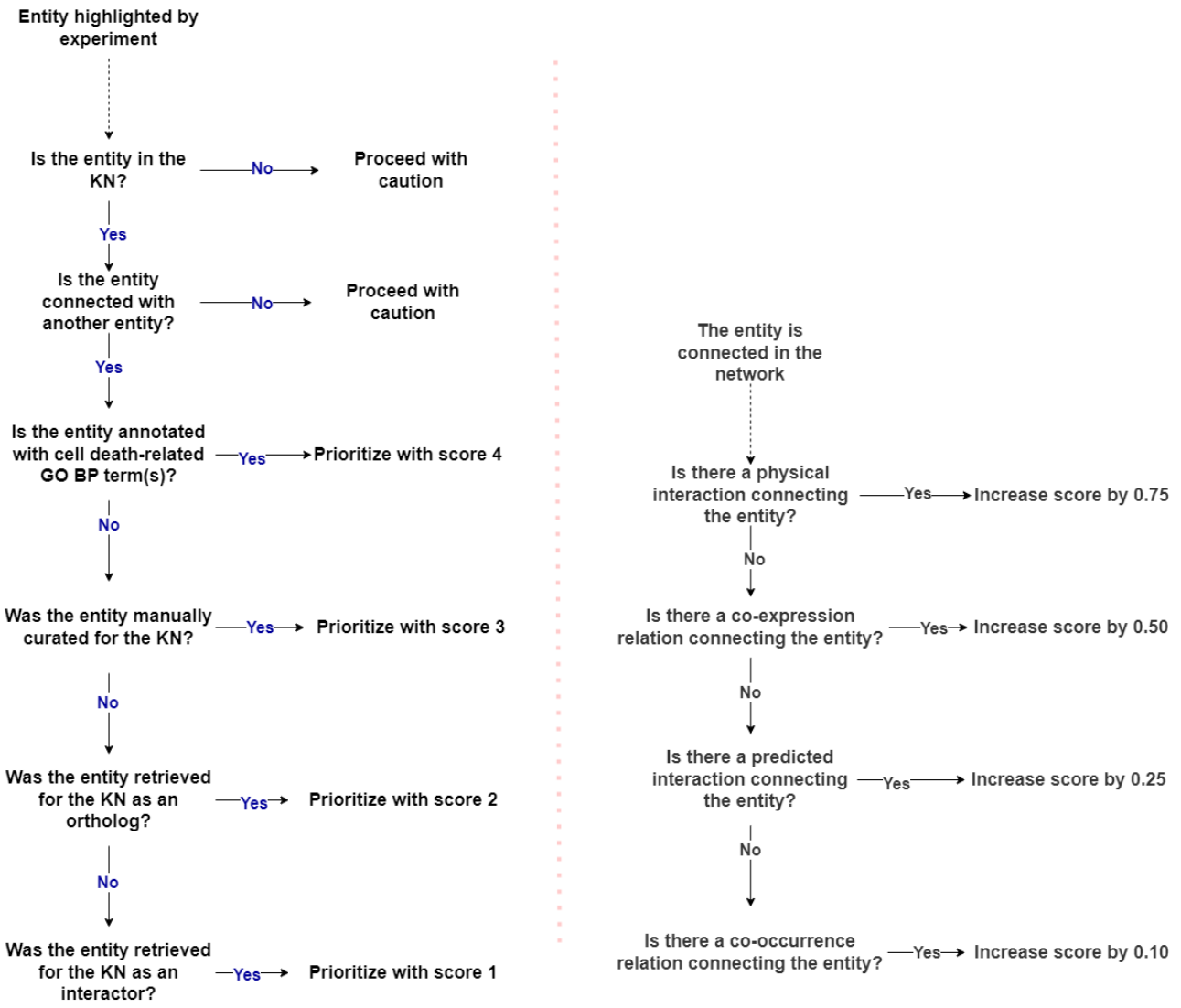



*Figure 15: A suggestion for how a decision tree may look like and how it might be used to prioritize new research by giving scores according to how likely an entity of interest is to be involved in plant RCD. The highest score signifies the highest likelihood. The ranking system starts (left side) to rank according to how the entity (if present) was retrieved and incorporated into the knowledge network (KN). Following, (right side) if the entity is connected in the KN the relation type can add to the score. "Proceed with caution" signifies that this entity (currently) is probably not worthwhile looking into.*

the KN. The latter can be changed with improvements to the methods and greater knowledge about the topic, while changes to the former depend on data availability and future research.

Even though the KN is generated with comparatively few inputs (few GO terms, species, and manually curated entities and interactions) it is large and complex. Someone might label the network as noisy. To answer specific questions there most often will be the need to interpret subsections of the KN rather than the whole. The KN was constructed so that the user has the flexibility to focus on specific entities and relations while simultaneously interpreting them with their connections to the extended neighborhood.

The KN can be considered to have fairly little specificity as it is a collective of entities from a multitude of RCD processes, described as occurring in various tissues or cells and induced by various conditions. A more extensive annotation process (with more fine-grained annotations) could have resolved the issue of grouping entities according to e.g. tissue, cellular localization, GO term annotations, etc. However, the current annotation state provided by the experts may not yet facilitate such grouping. For instance, as was shown for the state of the cell death-related GO BP term annotations in *A. thaliana* (i.e. figure 4a), other than annotations of "plant-type

hypersensitive response" (GO:0009626), most other annotations are of higher-order cell death-related GO BP terms. Manual annotation could be an option, although this is time-consuming and requires knowledge (on e.g. experimental methods) surpassing what I currently possess. One could resolve this issue with specificity by generating KN on the basis of more specific criteria. For instance, utilizing lower-order GO terms, and manual curation of only entities that are described as involved in specific cells or scenarios. However, it might be that the specificity of the KN should remain broad, the issue of specificity can rather be handled when extracting information that will be incorporated into PD maps.

## 8.3 Discussion of the data retrieval

### 8.3.1 Available information

The quality of the KN is inherently limited by the data available from the resources used to build it. It is reasonable to assume that much of the biological information is present to a large extent solely in the original papers rather than in biological databases. An argument for this would be that manual curation is time-consuming in itself, in addition to needing domain-specific knowledge, which can be scarce.

Some self-criticism must be accepted regarding the process of manually curating entities and relations for this project. There is undoubtedly more readily data available in research articles that should have been interpreted more extensively. As already mentioned, such a process requires more time and knowledge than what I was able to gather for this Master's thesis project. In an alternative scenario where the manual curation would have been performed more extensively, and with better quality, the KN is likely to have been of higher quality. These thoughts are addressed to encourage others with more domain-specific knowledge to utilize the developed data pipeline to generate new versions of the KN.

### 8.3.2 Curation criteria

The criteria used to curate the data will impact the quantity and quality of data included. Where to set the threshold needs to consider the loss of potentially important data and the inclusion of overwhelming noise. In this project, the noise is arguably known, or predicted, data (i.e. entities or interactions) that is not particularly connected to processes of RCD. At the same time, it is difficult to determine if this connection is either not occurring *in vivo*, has not been discovered, or is not yet accessible in databases. Whatever the source of the noise, a user will have to deal with it when using the knowledge network.

In this project, there were efforts to limit the number of entities that were incorporated as part of the KN. One of these efforts was to limit the number of proteins that were included on the basis of having experimentally validated physical interactions with other proteins that were already in the KN. There would be an option to iteratively add more entities to the KN, on the basis that they had interactions with the entities currently in the KN. However, doing so would undoubtedly contribute to increased noise, as the interaction could be even less likely to be relevant for RCD in *A. thaliana*. Figure 16 showcases an attempt to explain these remarks graphically.

**Gene Ontology**

The GO was utilized in this project by retrieving all annotations of a high-order GO BP term ("cell death" (GO:0008219)) or descendants of this term. By choosing higher-order GO BP terms this method is largely inclusive as it will retrieve annotations with all descendant terms as well. For this project, only the mentioned GO BP term was used to find annotated entities. This term was used because the annotated entities were all expected to be involved in core processes of RCD, and not only a part of a process that in certain instances can lead to RCD. Other GO BP terms could have been suitable for the generation of a KN of RCD. Examples of potentially suitable GO BP terms are many of the child terms of "response to stress" (GO:0006950). These include responses

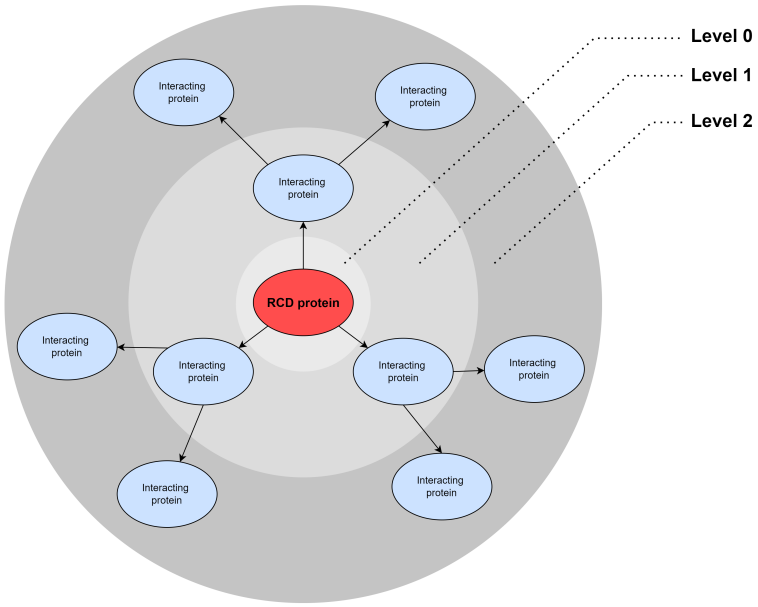



*Figure 16: A sketch of the iterative process of including ever more proteins to the knowledge network due to the proteins having experimentally validated interactions with proteins already in the network. Level 0 contains a protein that is classified as taking part in RCD, level 1 contains proteins that have an interaction with the RCD protein, while level 2 contains proteins that have an interaction with proteins of level 1. In this project, only proteins up to level 1 were included in the network.*

to salt, hypoxia, heat, and cold. Other could be the GO terms defense responses and senescence. As many processes can lead to RCD in plants there are likely many proteins that could be assigned as having a role in RCD.

The GO annotations with evidence groups IEA and ND were excluded in the process of generating the KN. The reason for this was first and foremost the idea that the first version of the KN would contain a smaller subset of proteins expertly annotated as entities of cell death processes. There is about a 38 times increase (from 410 to 15 770) in the number of *Viridiplantae* proteins annotated with a cell death-related GO BP term when including the automatically annotated. When only looking at *A. thalina* annotated proteins the increase is about twofold (104 to 225). Although this was chosen for the generation of the presented KN, the data pipeline used for the generation is easily tunable to facilitate the inclusion of annotations with any evidence. No cell death-related GO BP term annotations had the evidence group ND, so the exclusion of this group had no effect. As the ND evidence is supposed to only be allocated to annotations with the root GO terms (Molecular Function, Cellular Component, and Biological Process) no annotations would be found since the terms searched for in this project are of lower order than the root terms.

Without knowing the full extent of how proteins are automatically annotated, I would expect them to be so on the basis of what is already expertly annotated. If that is the case, one would require annotations with evidence of high quality for the automatic annotation to be performed accurately. As shown in the results there are a fair amount of annotations of cell death-related GO BP terms in *A. thaliana* that were annotated by experts (127 total). Moreover, most of these were annotated on the evidence of mutant phenotype (78% of 127 annotations), which can be considered evidence with high quality. However, many of these annotations are with higher-order terms. If methods of automatic annotation hinge on expert annotation of higher-order GO terms, I would expect the criteria for automatic annotation to be equally broad. One could imagine that this could be the reason why proteins have by automatic methods been annotated with GO BP terms containing the word "apoptosis", with apoptosis being an unsuitable term to use in the description of plant RCD [9, 10].

Another aspect regarding the GO annotations was the consideration of which relationship types between the GO terms themselves that were accepted. The relationship criteria used when searching for GO BP annotations were

"is a", "part of", "occurs in", which are the default relationship settings for QuickGO. The GO apparently has other relationships between terms, like "Regulates", "Positively regulates", and "Negatively regulates", which are not presented as options in the web resource of QuickGO. These relationships were therefore not considered. However, by manually adding the relationship "Regulates" as a parameter value in the web URL there was an increase in the number of annotations. This method produced an error when including the relationship types "Positively regulates", and "Negatively regulates" in the URL, so these were not looked further into. With the GO version 2024-05-24 (i.e. a later version than the one used when generating the KN) adding "Regulates" increased the number of *A. thaliana* cell death-related GO BP term annotations from 298 to 483 when all annotation evidence was accepted. When only considering annotations performed by experts the annotation number increased from 133 to 207. According to these numbers I have missed many proteins that may be important in the regulation of RCD in *A. thaliana*.

### *A. thaliana* orthologs of potential RCD entities in other plant species

There is likely information on RCD proteins in other plant species that were not accessed with the methods used in this project. For instance, the literature assessed in the process of manually curating *A. thaliana* RCD proteins also contained information on proteins of other plant species. One would hope that the information on the proteins of the other species was annotated to the proteins (e.g. as cell death-related GO BP terms). If so, the methods of this project would have included *A. thaliana* orthologs (if any) of these proteins. However, to be sure that these were included in the KN one could manually look for and include these orthologs in the KN.

There are other putative interactions that should have been included in the KN. If there is an experimentally validated interaction between the proteins of interest in another species, there may also be an interaction between the *A. thaliana* orthologs. Figure 17 aims to present this graphically. The inclusion of these putative interactions was discussed but not implemented in this version of the KN. The interactions found between the RCD proteins of other species could have been retrieved from IntAct or from databases where the evidence for a presented interaction is particularly strong, such as Reactome for human proteins.

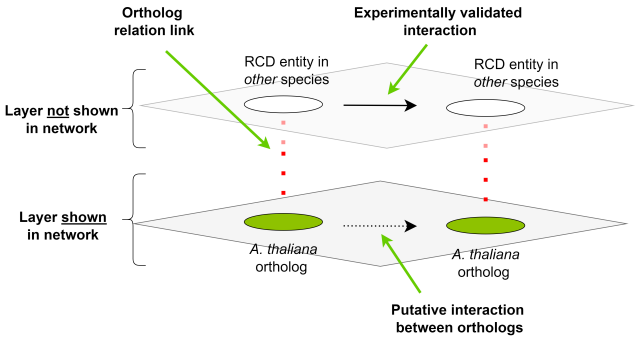



*Figure 17: Example showing how a validated interaction in one species can be a putative interaction between orthologs identified in another species. In this figure, green circles signify A. thaliana orthologs of RCD proteins identified in the other species (i.e. the white circles). The RCD entities in the other have an experimentally validated interaction (solid-line arrow), which can be a putative interaction between the respective A. thaliana orthologs (dashed arrow).*

### IntAct

With the settings chosen for this project, new entities from IntAct were imported to the KN regardless of the MIscore (i.e. IntAct's evidence score) for the interactions. This meant the inclusion of many interacting entities where the interaction evidence could be considered weak. Weak here refers to one or only a few experiments, where they test if there is an interaction between a substantial number of proteins, in a setting not necessarily similar to *in planta* conditions, and with few or a single experimental detection method. The effect of this decision can be seen in the KN (figure 6) where certain nodes are connected to many nodes in a star-like configuration, with interactions having a comparatively low MIscore (<0.40). It must be pointed out, that an entity can naturally have many interactions, e.g. if the entity is a transcription factor (e.g. NAC089

(UniProt:Q94F58)) or regulator of many proteins (e.g. calmodulin). In this project, I erred on the side of inclusion by including all interactions since having a threshold could result in possibly important interactions for RCD falling under the threshold. Moreover, having no retrieval threshold gives the user the possibility to filter the results themselves. What is preferable depends on the user's task at hand, e.g. finding highly validated interactions is useful for the generation of process diagrams, while identifying interactions with weak evidence is beneficial when determining what needs to be tested further.

**Co-expression**

Gene expression is a highly regulated process, where the combination of present regulators will have an impact on the expression levels. With combinatorial gene regulation in mind, I would be hesitant to accept that genes are always co-expressed regardless of the environmental conditions. In this project, co-expression relations were retrieved from GeneMANIA (and STRING but without references to the original study) regardless of the original studies and under what conditions the co-expressions were identified. As there probably are transcriptomic data available that were acquired under RCD-induced conditions, I could have used this data to identify co-expression relations specifically under RCD conditions. I would then not have to could then have these as the sole co-expression relations in the KN.

### 8.3.3 Biological identifiers

Choosing the type of biological identifiers can be challenging, as there are a plethora of them. A highly utilized identifier in the scientific literature for *A. thaliana* genes is the gene locus code (AGI (Arabidopsis Genome Initiative) locus code). It is beneficial as the gene locus is fixed, leading to no reason to change the code over time. In contrast, gene names (a.k.a. gene symbols) tend to be altered. Although it is common to refer to a gene identifier, it can have some drawbacks. For instance, when attributing annotations to a gene the annotations need to also hold true for all gene products of this gene since they all share the same gene identifier. When referring to a specific gene product I suggest using the UniProt accession code. It is a valid identifier for many biological databases, and even though it may be subject to change UniProt redirects the user to the most up-to-date accession code and entry page if an outdated identifier is used in a search query.

In this project, BioMart was used to convert gene identifiers to UniProt accession codes. In certain cases, the conversion from gene to gene product identifiers posed some problems. For instance, for some gene identifiers, BioMart was unable to find the UniProt accession code(s) or assigned them to the wrong gene. Overall, this issue was deemed to have fairly little impact on the quality of the KN. However, it underlines the importance of accurately referring to genes and gene products in the literature, so that data can correctly be attributed to the given biological entity.

### 8.3.4 The use of APIs in data retrieval

The benefit of using the APIs provided by biological databases is that one can retrieve the most up-to-date information by programmatic means. This process is so much faster than manually retrieving information for every single database entry of interest. However, even with the use of APIs, the data retrieval process can take time if data is retrieved for a large number of database entries. It varies depending on the database and what is retrieved, but in my experience the retrieval time per entry is up to a second. This will add up to 17 minutes if data from 1 000 database entries are retrieved. Some databases even enforce a maximum rate of requests for data retrieval (e.g. one entry per second), to prevent users from taking up too much of the service. The retrieval time also depends on the size of the data. There are available options to filter the data before it is retrieved to your local project (e.g. SPARQL queries), which may be desirable for future development of the data pipeline. Another unfortunate problem that became apparent during the development of the data pipeline, was that many APIs have a limit on how many entities can be part of a search query. An alternative to the use of APIs could be to use Cytoscape applications developed to directly within Cytoscape retrieve and incorporate

the data from certain databases into networks. IntAct and GeneMANIA are examples of databases that have such Cytoscape applications. The retrieval methods used by these applications are probably highly optimized. However, these were not used because I wanted to process the data as part of the data pipeline to specifically fit the format of the KN here generated.

## 8.4 Future projects

### 8.4.1 High-quality process diagrams

As described in the introduction, models describing as much information as possible regarding the workings of a biological process are of great value. However, generating these models requires great expertise in the field of study. Conceptual maps should be created using the SBML format, preferably in the process description (PD) format, as it most accurately describes the workings of the system. Producing high-quality PD maps is time-consuming because the quantity of data required to describe a biological process is large. The KN, being a collective of data, can be a resource when producing PD maps. The retrieved physical interactions can potentially be directly transferred to PD maps if the evidence is sufficient (which can be inferred from e.g. the confidence value (i.e. the IntAct MIscore)). The KN may also be used to get clues on which entities to search for in the literature, where there might be enough information to incorporate the entities into the PD maps.

As mentioned earlier, Burke et. al. concluded that only a small set of genes were DEGs in all experiments where RCD was induced by three different methods (salicylic acid, heat shock, and critical dilution culture) [39]. This observation can lead to the hypothesis that potentially very few genes are regulated similarly across different plant RCD processes. In that case, it may also be that the RCD processes are facilitated by different groups of entities. If so, it might be more fruitful to generate separate PD maps for the separate RCD processes, rather than producing a collective map for all. In the process of making these PD maps it might be useful to consult a KN that is a bit more specific to the process in question than the KN presented in this project. A future project could therefore be to generate such new KNs based on other GO BP terms and only manually curated entities that have been specifically described as being part of the process in question.

Future projects that extend upon the project presented in this thesis should draw inspiration from the newly available Stress Knowledge Map (SKM). The SKM encompasses two distinct networks serving their respective purposes. The Plant Stress Signalling model is a conceptual mechanistic model of the stress response cascade found in plant cells. This model is formulated to have a clear process description of the system, which would be desirable to describe processes observed in plant RCD. As stress can induce RCD in plants it would be expected that the Plant Stress Signalling model covers entities and reactions that would involved in RCD. This means that in the endeavors of generating PD maps for RCD, one might want to extend upon the Plant Stress Signalling model. The other network that is part of the SKM is referred to as the comprehensive Knowledge Network. As with the knowledge network generated by me, the SKM comprehensive Knowledge Network consists of a large collection of knowledge that can be valuable in the generation of PD maps or dynamical models. The comprehensive Knowledge Network is much larger than my knowledge network, meaning that there are likely aspects of the generation methods that should be evaluated for future development of the data pipeline used in this project [47].

### 8.4.2 Increased data retrieval

Today there is a lot of information freely available for anyone who knows how to put the data to good use. One therefore needs to have the knowledge of where to find it, how to analyze it, what it means, and how to connect all the information together. The biological databases simplify many of these processes. The resources used in this project are in many aspects considered state-of-the-art databases, and there are many more that were not utilized. The following section will showcase some of the quality resources that were not used in this project

due to time constraints and lack of knowledge about their qualities.

BioGRID (thebiogrid.org) is an example of a database that stores physical interaction data, similar to IntAct. From BioGRID should therefore be taken any interaction that is not already retrieved from IntAct. A physical interaction type that was not retrieved for this project was enzymatic reactions (MI:0414). This interaction type can be acquired from the Rhea (rhea-db.org) database. This information would be valuable to bridge the gap between proteins and other chemical substances in the KN, e.g. as seen for entities involved in ROS production and degradation. Moreover, Rhea also stores data on so-called transporter reactions. Information on these transportation events can be highly useful to understand for instance the workings of how $Ca^{2+}$ transport affects RCD. In Rhea, the chemical substances are identified using the ChEBI database identifier. In this project I used PubChem for annotation data on chemical substances, however, ChEBI could have served the same purpose. Moreover, as ChEBI, along with most other databases utilized in this project, is part of the ELIXIR core data resource, for continuity it could have been better to keep to ELIXIR databases.

There was an effort to utilize BAR's Arabidopsis Interaction Viewer in this project. However, the web application could not handle the number of genes that were searched for without crashing. The benefit of this resource was that it gave (in addition to PPIs) protein-DNA interactions which are not currently present in the KN. Protein-DNA interactions were retrieved from IntAct, but were removed in the data processing steps between retrieval and incorporation in the KN because the genes had a non-UniProt identifier. The latter was the Ensemble gene identifier. In hindsight, the issue with this identifier could have been resolved with ID mapping using BioMart to get the preferred UniProt accession code.

### 8.4.3 Filtering references on keywords

Through automated processes, one could limit the amount of nodes and links in the knowledge network by filtering on keywords in the references. One could filter on keywords in publicly available abstracts (e.g. from PubMed). Alternatively, one could use the search tools that can check if keywords are in the text. Such keywords could be "cell death", "plant", "Arabidopsis", etc. The keywords need to be words, or a collection of words, that should undoubtedly be in a text on the topic of interest. If the article does not show as one of the search entries it will not contain the keywords and will likely not describe the topic of interest.

### 8.4.4 Improvements to the code

**Logging**

The traceability of the data would be improved by incorporating logging into the code. The logging file(s) will present metadata to the user, e.g. the version number of the database and the exact moment the data was acquired from the given database. Moreover, logging can give feedback on potential issues, like faulty requests to the database (i.e. an error using the database API) that could be fixed with some manual intervention.

**Utilizing the biomaRt R package to map identifiers automatically**

The Ensembl tool BioMart is accessible through the R programming language package known as biomaRt. Unfortunately, this package does not have a Python equivalent. This means that to access the BioMart tool through other means than the web application we would need a separate R script. The inclusion of this R script in the data pipeline developed for this project would further decrease the number of manual interventions needed in the generation of new KNs. Another way of solving this problem, without turning to another programming language, would be to use another identifier mapping tool. UniProt provides an in-house ID mapping tool, which is accessible through an API.

# 9 Conclusion

From the project presented in this thesis can be drawn some conclusions that should be of value for future projects. The literature review revealed that most articles start by acknowledging that many of the mechanisms that facilitate cell death in a regulated fashion are unknown. My experience from the literature review was that it is difficult to keep track of all the entities and how they interact and regulate each other to facilitate RCD. This highlights the need for structural conceptual models that can, firstly, be assessed when trying to understand the system, and secondly, expanded upon when experimental findings indicate new knowledge.

Assessing the state of protein GO annotations of cell death-related BP terms revealed that most annotations are automatically attributed, both for all species in the *Viridiplantae* clade and more specifically for *A. thaliana*. The comparatively few annotations that are attributed by experts were either of the term "plant-type hypersensitive response", or the higher-order terms "cell death" or "programmed cell death". This highlights the need for annotation of terms that signify more specific RCD processes, especially RCD processes other than hypersensitive response (HR). On the basis that most of the annotations attributed to *A. thaliana* proteins is based upon evidence of an observed phenotype in mutants, it can also be concluded that more experimental evidence is needed to explain how entities of RCD affect one another (i.e. how they interact).

On the basis of orthology were identified *A. proteins* orthologs of proteins in other *Viridiplantae* species or *H. sapiens* that were expertly annotated with cell death-related GO BP terms. As these orthologs may have conserved functional properties the *A. thaliana* orthologs may be involved in RCD. Over 400 *A. thaliana* orthologs of annotated human proteins were identified. Many of these orthologs had connections with other proteins in the KN, while others had none. Only 35 *A. thaliana* orthologs were identified for other *Viridiplantae* species.

The knowledge network (KN) generated in this project can help in the endeavors of generating conceptual models. This is because it can help target more connected entities. The relations are of various types (from functional like physical interactions or co-expression, or simply connections of co-occurrences in the scientific literature), meaning that they signify various knowledge, and therefore must be interpreted accordingly. In the KN, the entities that had the highest number of connections tended to be entities considered to be involved in RCD, i.e. those that were either manually curated or retrieved from databases due to being annotated with cell death-related GO BP terms.

The KN can be used to evaluate and prioritize new research findings according to how likely it is that an entity is involved in plant RCD. This was shown with three genes that were present in the KN and identified as DEGs under heat shock-induced RCD conditions in a study.

# References


[1] FAO, IFAD, UNICEF, WFP, WHO. In Brief to The State of Food Security and Nutrition in the World 2023. Rome, Italy: FAO; 2023. Available from: https://openknowledge.fao.org/handle/20.500.14283/cc6550en.

[2] Kacprzyk J, Burke R, Armengot L, Coppola M, Tattrie SB, Vahldick H, et al. Roadmap for the next decade of plant programmed cell death research. New Phytologist. 2024;242(5):1865-75. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.19709. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.19709.

[3] Dauphinee AN, Gunawardena AN. An Overview of Programmed Cell Death Research: From Canonical to Emerging Model Species. In: Gunawardena AN, McCabe PF, editors. Plant Programmed Cell Death. Cham: Springer International Publishing; 2015. p. 1-31. Available from: https://doi.org/10.1007/978-3-319-21033-9_1.

[4] Locato V, De Gara L. Programmed Cell Death in Plants: An Overview. In: De Gara L, Locato V, editors. Plant Programmed Cell Death: Methods and Protocols. New York, NY: Springer; 2018. p. 1-8. Available from: https://doi.org/10.1007/978-1-4939-7668-3_1.

[5] Galluzzi L, Vitale I, Aaronson SA, Abrams JM, Adam D, Agostinis P, et al. Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018. Cell Death & Differentiation. 2018 Mar;25(3):486-541. Number: 3 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41418-017-0012-4.

[6] Aguilera A, Distéfano A, Jauzein C, Correa-Aragunde N, Martinez D, Martin MV, et al. Do photosynthetic cells communicate with each other during cell death? From cyanobacteria to vascular plants. Journal of Experimental Botany. 2022 Dec;73(22):7219-42. Available from: https://doi.org/10.1093/jxb/erac363.

[7] Daneva A, Gao Z, Van Durme M, Nowack MK. Functions and Regulation of Programmed Cell Death in Plant Development. Annual Review of Cell and Developmental Biology. 2016;32(1):441-68. Available from: https://doi.org/10.1146/annurev-cellbio-111315-124915.

[8] Huysmans M, Lema A S, Coll NS, Nowack MK. Dying two deaths — programmed cell death regulation in development and disease. Current Opinion in Plant Biology. 2017 Feb;35:37-44. Available from: https://www.sciencedirect.com/science/article/pii/S1369526616301923.

[9] van Doorn WG, Beers EP, Dangl JL, Franklin-Tong VE, Gallois P, Hara-Nishimura I, et al. Morphological classification of plant cell deaths. Cell Death & Differentiation. 2011 Aug;18(8):1241-6. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/cdd201136.

[10] Minina EA, Dauphinee AN, Ballhaus F, Gogvadze V, Smertenko AP, Bozhkov PV. Apoptosis is not conserved in plants as revealed by critical examination of a model for plant apoptosis-like cell death. BMC Biology. 2021 May;19(1):100. Available from: https://doi.org/10.1186/s12915-021-01018-z.

[11] Mazein A, Acencio ML, Balaur I, Rougny A, Welter D, Niarakis A, et al. A guide for developing comprehensive systems biology maps of disease mechanisms: planning, construction and maintenance. Frontiers in Bioinformatics. 2023 Jun;3. Publisher: Frontiers. Available from: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1197310.

[12] Novère NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, et al. The Systems Biology Graphical Notation. Nature Biotechnology. 2009 Aug;27(8):735-41. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nbt.1558.

[13] Hastings J. Primer on Ontologies. In: Dessimoz C, Škunca N, editors. The Gene Ontology Handbook. New York, NY: Springer; 2017. p. 3-13. Available from: https://doi.org/10.1007/978-1-4939-3743-1_1.

[14] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000 May;25(1):25-9. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ng0500_25.

[15] The Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023 May;224(1):iyad031. Available from: https://doi.org/10.1093/genetics/iyad031.

[16] Nadendla S, Jackson R, Munro J, Quaglia F, Mészáros B, Olley D, et al. ECO: the Evidence and Conclusion Ontology, an update for 2022. Nucleic Acids Research. 2022 Jan;50(D1):D1515-21. Available from: https://doi.org/10.1093/nar/gkab1025.

[17] Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, et al. The HUPO PSI's Molecular Interaction format—a community standard for the representation of protein interaction data. Nature Biotechnology. 2004 Feb;22(2):177-83. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nbt926.

[18] Koonin EV. Orthologs, Paralogs, and Evolutionary Genomics1. Annual Review of Genetics. 2005 Dec;39(Volume 39, 2005):309-38. Publisher: Annual Reviews. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725.

[19] Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research. 2013 Jan;41(D1):D808-15. Available from: https://doi.org/10.1093/nar/gks1094.

[20] Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research. 2023 Jan;51(D1):D638-46.

[21] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research. 2023 Jan;51(D1):D523-31. Available from: https://doi.org/10.1093/nar/gkac1052.

[22] Villaveces JM, Jiménez RC, Porras P, del Toro N, Duesbury M, Dumousseau M, et al. Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. Database. 2015 Jan;2015:bau131. Available from: https://doi.org/10.1093/database/bau131.

[23] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research. 2014 Jan;42(D1):D358-63. Available from: https://doi.org/10.1093/nar/gkt1115.

[24] Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics. 2009 Nov;25(22):3045-6. Available from: https://doi.org/10.

1093/bioinformatics/btp536.

[25] Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva E, et al. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Research. 2023 Jan;51(D1):D445-51. Available from: https://academic.oup.com/nar/article/51/D1/D445/6814468.

[26] Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar G, et al. InterPro in 2022. Nucleic Acids Research. 2023 Jan;51(D1):D418-27. Available from: https://doi.org/10.1093/nar/gkac993.

[27] Help · GeneMANIA;. Available from: https://pages.genemania.org/help/.

[28] Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. genesis. 2015;53(8):474-85. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/dvg.22877.

[29] Harrison PW, Amode MR, Austine-Orimoloye O, Azov A, Barba M, Barnes I, et al. Ensembl 2024. Nucleic Acids Research. 2024 Jan;52(D1):D891-9. Available from: https://doi.org/10.1093/nar/gkad1049.

[30] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Research. 2023 Jan;51(D1):D1373-80. Available from: https://doi.org/10.1093/nar/gkac956.

[31] yWorks GmbH. yEd; 2024. Available from: https://www.yworks.com/products/yed.

[32] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003 Nov;13(11):2498-504. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Available from: https://genome.cshlp.org/content/13/11/2498.

[33] pandas development team T. pandas-dev/pandas: Pandas. Zenodo; 2023. Available from: https://zenodo.org/records/8092754.

[34] Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007 May;9(3):90-5. Conference Name: Computing in Science & Engineering. Available from: https://ieeexplore.ieee.org/document/4160265.

[35] Inc PT. Collaborative data science. Montreal, QC: Plotly Technologies Inc.; 2015. Available from: https://plot.ly.

[36] Inferred from Biological aspect of Ancestor (IBA) - GO Wiki;. Available from: https://wiki.geneontology.org/index.php/Inferred_from_Biological_aspect_of_Ancestor_(IBA).

[37] Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Briefings in Bioinformatics. 2011 Sep;12(5):449-62. Available from: https://doi.org/10.1093/bib/bbr042.

[38] Tintor N, Ross A, Kanehara K, Yamada K, Fan L, Kemmerling B, et al. Layered pattern receptor signaling via ethylene and endogenous elicitor peptides during Arabidopsis immunity to bacterial infection.

Proceedings of the National Academy of Sciences. 2013 Apr;110(15):6211-6. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1216780110.

[39] Burke R, McCabe A, Sonawane NR, Rathod MH, Whelan CV, McCabe PF, et al. Arabidopsis cell suspension culture and RNA sequencing reveal regulatory networks underlying plant-programmed cell death. The Plant Journal. 2023;115(6):1465-85. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.16407.

[40] Herrera-Ubaldo H, Campos SE, López-Gómez P, Luna-García V, Zúñiga-Mayo VM, Armas-Caballero GE, et al. The protein–protein interaction landscape of transcription factors during gynoecium development in Arabidopsis. Molecular Plant. 2023 Jan;16(1):260-78. Available from: https://www.sciencedirect.com/science/article/pii/S1674205222002982.

[41] Wanamaker SA, Garza RM, MacWilliams A, Nery JR, Bartlett A, Castanon R, et al. CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. Nature Methods. 2017 Aug;14(8):819-25. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nmeth.4343.

[42] Chang IF, Curran A, Woolsey R, Quilici D, Cushman JC, Mittler R, et al. Proteomic profiling of tandem affinity purified 14-3-3 protein complexes in Arabidopsis thaliana. PROTEOMICS. 2009;9(11):2967-85. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200800445.

[43] Zuber H, Davidian JC, Aubert G, Aimé D, Belghazi M, Lugan R, et al. The seed composition of Arabidopsis mutants for the group 3 sulfate transporters indicates a role in sulfate translocation within developing seeds. Plant Physiology. 2010 Oct;154(2):913-26.

[44] Thomas EL, Van der Hoorn RAL. Ten Prominent Host Proteases in Plant-Pathogen Interactions. International Journal of Molecular Sciences. 2018 Feb;19(2):639. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/1422-0067/19/2/639.

[45] Hara-Nishimura I, Hatsugai N. The role of vacuole in plant cell death. Cell Death & Differentiation. 2011 Aug;18(8):1298-304. Number: 8 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/cdd201170.

[46] Lam E. Vacuolar proteases livening up programmed cell death. Trends in Cell Biology. 2005 Mar;15(3):124-7.

[47] Bleker C, Ramšak Z, Bittner A, Podpečan V, Zagorščak M, Wurzinger B, et al. Stress Knowledge Map: A knowledge graph resource for systems biology analysis of plant stress responses. Plant Communications. 2024 Apr:100920. Available from: https://www.sciencedirect.com/science/article/pii/S2590346224001901.

# Appendix

## A Automatic curation criteria

### A.1 GO query details

The parameter values for the search query used in QuickGO to retrieve all proteins of *Viridiplantae* annotated with cell death-related GO Biological Process terms (i.e. GO:0008219, or descendant terms) are presented in table 2.

*Table 2: The parameter values chosen for the search query to retrieve proteins of Viridiplantae (taxonomic ID: 33090) annotated with cell death-related GO Biological Process terms (i.e. GO:0008219, or descendant terms). The evidence codes signify all evidence groups except IEA and ND, which have the respective evidence codes ECO:0000501 and ECO:0000307.*

| Parameter | Value |
|---|---|
| goUsage | descendants |
| goUsageRelationships | is_a,part_of,occurs_in |
| goId | GO:0008219 |
| taxonId | 33090 |
| taxonUsage | descendants |
| geneProductType | protein |
| aspect | biological_process |
| evidenceCode | ECO:0000352,ECO:0000269,ECO:0000314, ECO:0000315,ECO:0000316,ECO:0000353, ECO:0000270,ECO:0007005,ECO:0007001, ECO:0007003,ECO:0007007,ECO:0006056, ECO:0000250,ECO:0000247,ECO:0000266, ECO:0000318,ECO:0000320,ECO:0000321, ECO:0000255,ECO:0000317,ECO:0000304, ECO:0000303,ECO:0000305,ECO:0000245 |

### A.2 IntAct search query details and handling of results

The query details for the IntAct API were the batch search as True, min MIscore as 0, max MIscore as 1, negative filter as "positive and negative", and the output format was set to "miTab26". These settings were believed to be the default settings for the batch search of the web application of IntAct.

Each observance of an interaction event between two (or more) entities is stored at IntAct as a separate IntAct entry, with an unambiguous IntAct accession code. The result of this is that each interaction can be described by multiple IntAct entries if multiple studies, or rather multiple experiments, have shown that the interaction occurs. The IntAct search returns individual IntAct entries, rather than individual interactions (e.g. a summary of all evidential data indicating an interaction). To get the results in a format where an interaction is described by a single row in a table, the information from each IntAct entry, of that interaction, was combined. The MIscore was used to identify if multiple IntAct entries represented the same interaction. This evidence value is calculated based on (among other parameters) the number of IntAct entries, i.e. it should be the same for all entries representing the same interaction. In summary, all IntAct entries having the same interacting partners and the same MIscore, were combined to represent all the evidence of an interaction.

### A.3 STRING query details

To retrieve entities and relations from the STRING database the "network" API was utilized. The query details were *A. thaliana* as the species, a required score set to 400 (which was the suggested default value), and the UniProt accession codes for the entities searched for. From the output was retrieved the co-expression and co-occurrence relations.

# B Supplementary methods

## B.1 BioMart queries

The settings chosen for the conversion of Gene stable IDs (also known as gene locus codes) and NCBI Gene IDs to UniProt accession codes in BioMart are respectively given in table 3 and 4. To get a 1:1 ratio of loaded to returned identifiers the result was filtered using a Python script with the logic shown in table 5.

*Table 3: BioMart query-settings for the conversion of Gene stable IDs (also known as gene locus codes) to UniProt accession codes. The operation was done as part of incorporating data from STRING into the knowledge network.*

| Option 1 | Option 2 | Option 3 |
|---|---|---|
| Database | Ensembl Plants Gene 58 | |
| Dataset | Arabidopsis thaliana genes (TAIR10) | |
| Filters | GENE | Input external references ID list → Gene stable ID |
| Attributes | GENE | Gene stable ID |
| | EXTERNAL | UniProtKB/Swiss-Prot ID<br>UniProtKB/TrEMBL ID |

*Table 4: BioMart query-settings for the conversion of NCBI IDs to UniProt accession codes. The operation was done as part of incorporating data from GeneMANIA into the knowledge network.*

| Option 1 | Option 2 | Option 3 |
|---|---|---|
| Database | Ensembl Plants Gene 58 | |
| Dataset | Arabidopsis thaliana genes (TAIR10) | |
| Filters | GENE | Input external references ID list → NCBI gene (formerly Entrezgene) ID(s) |
| Attributes | EXTERNAL | NCBI gene (formerly Entrezgene) ID<br>UniProtKB/Swiss-Prot ID<br>UniProtKB/TrEMBL ID |

*Table 5: The method used to get a single preferred identifier per inputted identifier in BioMart.*

| Filter level | Rule |
|---|---|
| 1 | Get the first available Swiss-Prot identifier |
| 2 | Get the first available TrEMBL identifier with 6 characters |
| 3 | Get the first available TrEMBL identifier |

## B.2 Specific methods used to generate results figures

1. Group nodes according to where identified (i.e. here; color)

2. Remove links between nodes within each group (in yEd: Tools → Select Elements → Edges → Select → "Selected Nodes Subgraph Edges" → Ok, followed by Delete)

3. Remove links between the group of manually curated nodes (i.e. red color) and the group of nodes annotated with cell death GO BP term(s) (in yEd: Select both groups, Tools → Select Elements → Edges → Select → "Selected Nodes Subgraph Edges" → Ok, followed by Delete)

4. Remove nodes with no links (in yEd: Tools → Select Elements → Nodes → Select → Degree → Ok, followed by Delete)

5. Organize the ortholog group so that one can distinguish between those nodes having a link to both a manually curated node and a node GO annotated, and those ortholog nodes having a link to either one or the other group.

    5.1 Select the ortholog group and another group, e.g. GO term annotated group, find the links between them, the nodes the links belongs to, and move the selection to separate the selection from the rest

of the nodes (in yEd: Select the two groups, followed by; Tools → Select Elements → Edges → Select → "Selected Nodes Subgraph Edges" → Ok, followed by; Tools → Select Elements → Nodes → Select → "Nodes of Selected Edges" → Ok, then move the selection).

5.2 Select the all nodes belonging to the ortholog group and the group not selected last time, i.e. the manually curated group. (in yEd: Select the two groups, followed by; Tools → Select Elements → Edges → Select → "Selected Nodes Subgraph Edges" → Ok, followed by; Tools → Select Elements → Nodes → Select → "Nodes of Selected Edges" → Ok, then move the selection).

6. Finally, organize the layout

# C Supplementary results

## C.1 GO terms

*Table 6: Cell death-related GO BP term annotations of Viridiplantae with all evidence groups*

| GO term | GO term name | Count | Percent |
|---|---|---|---|
| GO:0009626 | plant-type hypersensitive response | 8965 | 46 |
| GO:0012501 | programmed cell death | 3698 | 19 |
| GO:0010343 | singlet oxygen-mediated programmed cell death | 1425 | 7 |
| GO:0008637 | apoptotic mitochondrial changes | 1367 | 7 |
| GO:0006915 | apoptotic process | 1185 | 6 |
| GO:0070059 | intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress | 960 | 5 |
| GO:0006309 | apoptotic DNA fragmentation | 624 | 3 |
| GO:0097192 | extrinsic apoptotic signaling pathway in absence of ligand | 488 | 3 |
| GO:0043653 | mitochondrial fragmentation involved in apoptotic process | 344 | 2 |
| GO:0010198 | synergid death | 46 | <1 |
| GO:0008630 | intrinsic apoptotic signaling pathway in response to DNA damage | 42 | <1 |
| GO:0010623 | programmed cell death involved in cell development | 30 | <1 |
| GO:0008219 | cell death | 29 | <1 |
| GO:1902445 | regulation of mitochondrial membrane permeability involved in programmed necrotic cell death | 22 | <1 |
| GO:0008625 | extrinsic apoptotic signaling pathway via death domain receptors | 21 | <1 |
| GO:0097190 | apoptotic signaling pathway | 20 | <1 |
| GO:0070782 | phosphatidylserine exposure on apoptotic cell surface | 9 | <1 |
| GO:1902742 | apoptotic process involved in development | 9 | <1 |
| GO:0097191 | extrinsic apoptotic signaling pathway | 7 | <1 |
| GO:0010421 | hydrogen peroxide-mediated programmed cell death | 6 | <1 |
| GO:0048102 | autophagic cell death | 5 | <1 |
| GO:0034050 | symbiont-induced defense-related programmed cell death | 4 | <1 |
| GO:0051402 | neuron apoptotic process | 4 | <1 |
| GO:0036462 | TRAIL-activated apoptotic signaling pathway | 4 | <1 |
| GO:0097468 | programmed cell death in response to reactive oxygen species | 3 | <1 |

*Table 7: Cell death-related GO BP term annotations of Viridiplantae with excluding evidence group IEA and ND.*

| GO term | GO term name | Count | Percent |
|---|---|---|---|
| GO:0009626 | plant-type hypersensitive response | 366 | 84 |
| GO:0008219 | cell death | 29 | 7 |
| GO:0012501 | programmed cell death | 17 | 4 |
| GO:0010623 | programmed cell death involved in cell development | 8 | 2 |
| GO:0010198 | synergid death | 5 | 1 |
| GO:0010343 | singlet oxygen-mediated programmed cell death | 5 | 1 |
| GO:0006309 | apoptotic DNA fragmentation | 2 | <1 |
| GO:0010421 | hydrogen peroxide-mediated programmed cell death | 2 | <1 |
| GO:0048102 | autophagic cell death | 1 | <1 |
| GO:0034050 | symbiont-induced defense-related programmed cell death | 1 | <1 |
| GO:0097468 | programmed cell death in response to reactive oxygen species | 1 | <1 |

*Table 8: Cell death-related GO BP term annotations of A. thaliana with excluding evidence group IEA and ND.*

| GO term | GO term name | Count | Percent |
|---|---|---|---|
| GO:0009626 | plant-type hypersensitive response | 60 | 47 |
| GO:0008219 | cell death | 29 | 23 |
| GO:0012501 | programmed cell death | 17 | 13 |
| GO:0010623 | programmed cell death involved in cell development | 6 | 5 |
| GO:0010198 | synergid death | 5 | 4 |
| GO:0010343 | singlet oxygen-mediated programmed cell death | 5 | 4 |
| GO:0010421 | hydrogen peroxide-mediated programmed cell death | 2 | 2 |
| GO:0048102 | autophagic cell death | 1 | 1 |
| GO:0034050 | symbiont-induced defense-related programmed cell death | 1 | 1 |
| GO:0097468 | programmed cell death in response to reactive oxygen species | 1 | 1 |

### C.2 Imported interactors with high degree

*Table 9: Interactor nodes (those nodes retrieved from IntAct) that have a high degree (above 5) in the subsection of the knowledge network that only consists of nodes connected by physical interaction links.*

| Gene name | Degree | UniProt |
|---|---|---|
| NHL3 | 11 | Q9FNH6 |
| | 10 | Q9C7D7 |
| IQD6 | 9 | O64852 |
| | 9 | Q9AST5 |
| | 8 | Q2NJD6 |
| HHP2 | 8 | Q84N34 |
| MQB2.1 | 7 | Q8GX94 |
| TCP15 | 7 | Q9C9L2 |
| RLK7 | 7 | F4I2N7-2 |
| ERECTA | 7 | Q42371 |
| UBC34 | 7 | Q9SHI7 |
| VPS60-1 | 6 | Q9LPN5 |
| CYP21-4 | 6 | Q9C835 |
| BRI1 | 6 | O22476 |
| BAM3 | 6 | O65440-2 |
| NHL6 | 6 | Q8LD98 |
| KNAT1 | 6 | P46639 |
| | 6 | Q8L9S0 |
| LRR-RLK | 6 | Q9SVG8 |
| | 6 | Q9SX96 |

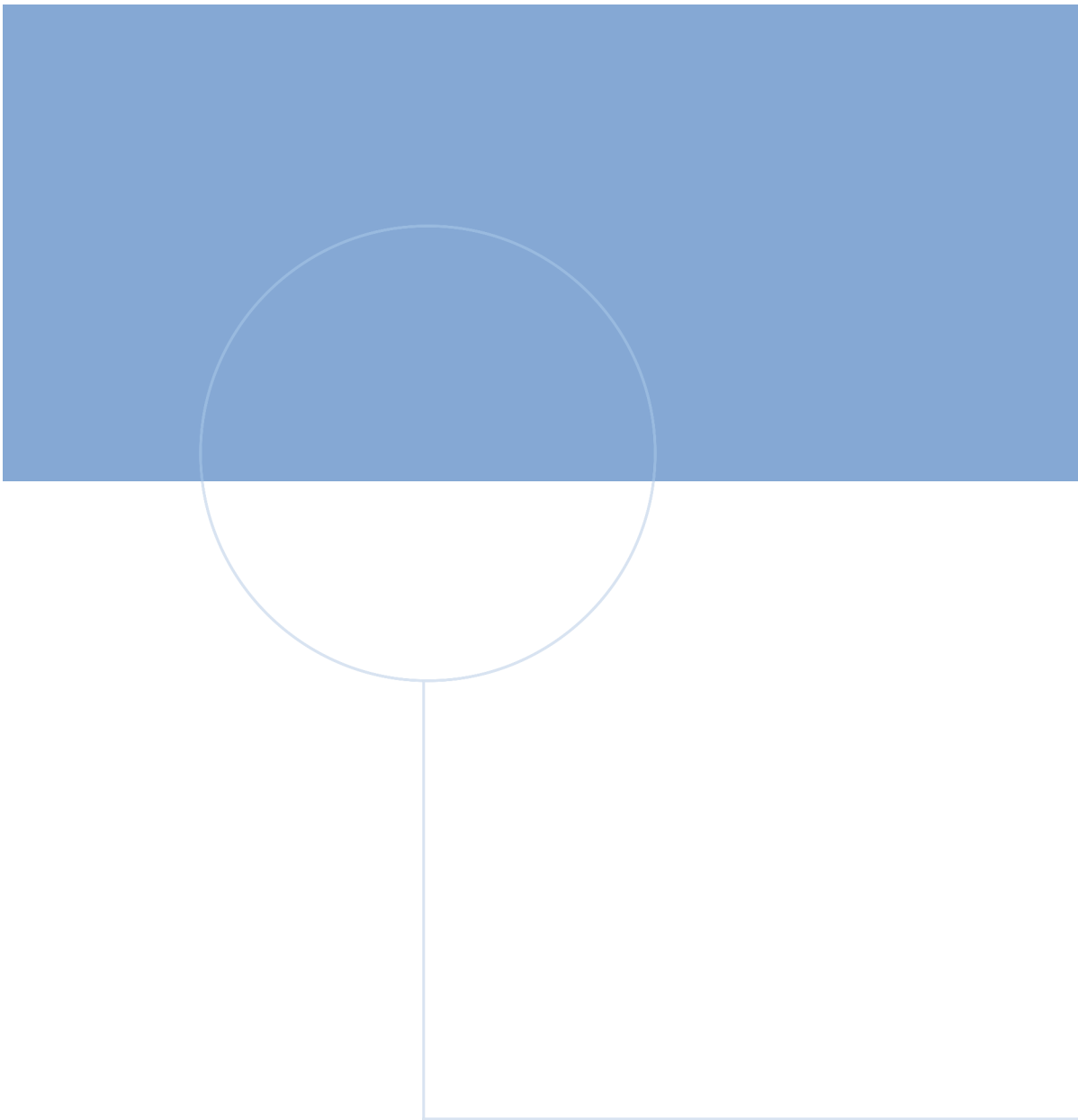
### C.3 Interesting nodes in regards to clustering coefficient

*Table 10: Information on proteins forming a 5-clique in the knowledge network. Many of the nodes that have a clustering coefficient of 1 form a clique with all or some of the proteins in this table. Retrieval methods are how the proteins were incorporated into the knowledge network. "Retrieved as GO term annotated" signifies that the node was included due to being annotated with a cell death-related GO BP term.*

| Gene name | UniProt accession | Degree | Clustering coefficient | Retrieval method |
|---|---|---|---|---|
| EFR | C0LGT6 | 98 | 0.10 | Manually curated |
| SERK4 | Q9SKG5 | 83 | 0.17 | Retrieved as GO term annotated |
| FLS2 | Q9FL28 | 123 | 0.12 | Manually curated |
| BIR1 | Q9ASS4 | 198 | 0.09 | Manually curated |
| BAK1 | Q94F62 | 218 | 0.20 | Manually curated |

*Table 11: Information on the nodes with the highest degree and clustering coefficient value of 1.00 in the knowledge network. All the nodes in the table formed 6-cliques with the nodes presented in table 10. Retrieval methods are how the proteins were incorporated into the knowledge network. "Retrieved as interactor" signifies that the protein was incorporated because it has an experimentally validated physical interaction with some other entity of the knowledge network.*

| Gene name | UniProt accession | Degree | Clustering coefficient | Retrieval method |
|---|---|---|---|---|
| ERL1 | C0LGW6 | 5 | 1.00 | Retrieved as interactor |
| ERL2 | Q6XAT2 | 5 | 1.00 | Retrieved as interactor |
| SERK5 | Q8LPS5 | 5 | 1.00 | Retrieved as interactor |
| SARK | Q8VYT3 | 5 | 1.00 | Retrieved as interactor |
| SRF6 | Q9C8M9 | 5 | 1.00 | Retrieved as interactor |
| NIK1 | Q9LFS4 | 5 | 1.00 | Retrieved as interactor |
| BRL3 | Q9LJF3 | 5 | 1.00 | Retrieved as interactor |
| LRR-RLK | Q9ZVD4 | 5 | 1.00 | Retrieved as interactor |



NTNU

Norwegian University of
Science and Technology