TITLE: PlantConnectome: knowledge networks encompassing >100,000 plant article abstracts

AUTHORS: Kevin Fo[1], Yu Song Chuah[1], Herman Foo[1], Emilia Emmanuelle Davey[1], Melissa Fullwood[1], Guillaume Thibault[1], Marek Mutwil[1*]

[1]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551

*Corresponding author:
Marek Mutwil
School of Biological Sciences,
Nanyang Technological University,60 Nanyang Drive,
Singapore 637551
Email: mutwil@ntu.edu.sg

Abstract
Predicting gene function is indispensable to our understanding of biology. However, these predictions hinge on large collections of experimentally characterized genes, the compilation of which is not only labor-intensive and time-consuming but rendered near-impossible given the volume and diversity of scientific literature. Here, we tackle this challenge by deploying the text-mining capacities of Generative Pre-trained Transformer (GPT) to process over 100,000 plant biology abstracts. Our approach unveiled nearly 400,000 functional relationships between a wide array of biological entities—genes, metabolites, tissues, and others—with a remarkable accuracy of over 85%. We encapsulated these findings in PlantConnectome, a user-friendly database, and demonstrated its diverse utility by providing insights into gene regulatory networks, protein-protein interactions, as well as developmental and stress responses. We believe that this innovative use of AI in the life sciences will significantly accelerate and direct research, drive powerful gene function prediction methods and help us keep up to date with the rapidly growing corpus of scientific literature.

Introduction
Gene function prediction is the keystone to our microscopic and macroscopic understanding of biology, revealing how genes contribute to the formation and mechanism of biological systems (Rhee and Mutwil, 2014) and providing insights into biological diversity and evolution (Yu et al., 2020; Guo et al., 2020). Besides verifying existing hypothetical connections between genes and their functions, gene predictions guide the identification of new gene-function relationships and elucidate evolutionary processes that shape biological mechanistic intricacies (Ruprecht et al., 2017; Julca et al., 2021). Moreover, predictive modeling can significantly refine experimental approaches, eliminating unnecessary tests on already-characterized genes and directing efforts toward the ones most likely to yield novel insights (Persson et al., 2005; Brown et al., 2005). Accordingly, numerous tools and databases providing gene function services have been developed, including STRING (Szklarczyk et al., 2015), GeneMANIA (Franz et al., 2018), CoNeKT (Proost and Mutwil, 2018), ATTED-II (Aoki et al., 2016), and others (Lim et al., 2022).

Predicting gene function requires two components: i) gene property data (e.g., coding sequence, expression patterns, and protein structure) and ii), gold standard data (i.e., genes with experimentally verified functions) (Rhee and Mutwil, 2014; Radivojac et al., 2013). The former is firstly used to connect uncharacterized genes with characterized ones similar in sequence or expression; based on the 'guilt-by-association' principle, the uncharacterized genes are subsequently labeled according to the functions of the characterized genes (i.e., the gold standard data) to which they were connected (Rhee and Mutwil, 2014).

Nonetheless, gene function prediction remains highly challenging due to the complexity and vastness of biological data, plateauing our understanding of plant genomes (Rhee and Mutwil, 2014) and, thus, our ability to address ever-exacerbating concerns in agriculture, medicine, and industry (National Research Council (US) Committee on Examination of Plant Science Research Programs in the United States, 1992). Specifically, establishing the gold standard necessitates manual, work-intensive extraction of gene functional information from scientific articles (Oughtred et al., 2021), preventing public repositories that harbor the gold standard data, such as BioGRID (protein-protein interactions, or PPIs) and AGRIS (gene regulatory networks, or GRNs)(Oughtred et al., 2021; Yilmaz et al., 2011), from keeping up to date with state-of-the-art knowledge. Furthermore, such repositories are typically restricted to specific data types (e.g., PPI or GRNs), precluding the integration of various data kinds that is critical to deepening our understanding of plant biology.

We, thus, seized the recent developments in Artificial Intelligence to revive this understanding plateau, deploying the advanced text mining capacities of a high-performance language model, Generative Pre-trained Transformer (GPT), to process over 100,000 research abstracts from leading journals in plant biology. Our approach excavated upwards of 300,000 functional relationships between more than 100,000 entities comprising genes, metabolites, tissues, organs, and other biological components. The manual inspection of these relationships revealed not only their impressive accuracy but exceptionally complementary insights, even doubling the amount of functional information relative to the current coverage of gene regulatory networks. Recognizing the potential of this data to enhance plant understanding, we constructed PlantConnectome, a user-friendly database comprising novel visuals that can illuminate gene function, organ development, gene regulatory networks, protein-protein interactions, and much more. PlantConnectome is available at the following URL: https://connectome.plant.tools/.

Materials and Methods

Retrieval of paper abstracts

Using BioPython version 1.81, we downloaded all abstracts published after 2005 from Plant Physiology, New Phytologist, the Journal of Experimental Botany, the Plant Journal, BioMed Central Plant Biology, Plant Cell, Plant Signal Behavior, Planta, Plant Cell Physiology, the Journal of Plant Physiology, Plant Cell Environment, Plant Molecular Biology, Physiol Plant, International Journal of Molecular Biology, Molecular Plant-Microbe Interactions, Molecular Plant, Proceedings of the National Academy of Sciences, Nature Plants, and the Journal of Integrated Plant Biology. Plant Science was one exception, from which we downloaded its post-2020 papers due to the relatively small number of gene functions that it captures (Figure S1). For each abstract, OpenAI's Python API for davinci 3.5 model was utilized as a part of a prompt (Table 1). The returned results were subsequently processed to remove single letter entities

(i.e., Gene !affects! X) and convert passive edges (i.e., Photosynthesis !is affected by! Sunlight) to active edges (Sunlight !affects! Photosynthesis), while edges with very similar meanings were grouped together and repesented by one edge. The model was run with default parameters, with the exception of temperature=0 to obtain deterministic results. In total, 101,341 abstracts were processed within two weeks.

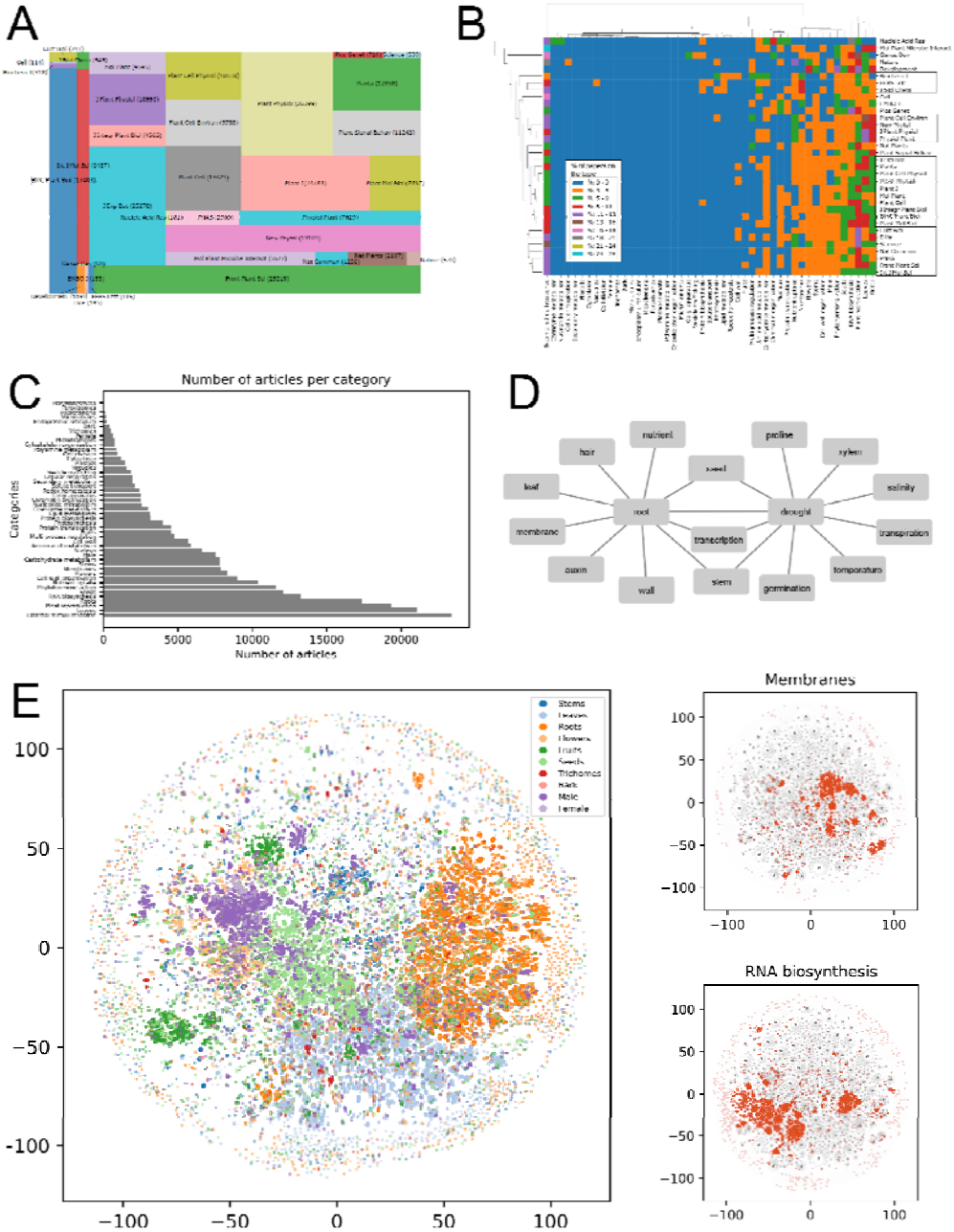

Figure 1. Meta-analysis of the article abstracts. A) Journal sources and the number of articles pertaining to plant research. B) Clustering of journals (rows), topics (columns), and the percentage of papers on a given topic per journal (cell color). C) The number of articles corresponding to the 49 major classes of topics. D) Co-occurrence network of keywords associated with 'root' and 'drought'. Nodes represent keywords, while edges connect keywords connected with the Jaccard Index value within the top 1% of the maximum JI values. E) t-SNE

visualization of the abstracts with a focus on plant organs (left panel), membranes (right, top panel), and RNA biosynthesis (right, bottom panel). Each point represents an article, and the colors indicate the different organs.

Construction of PlantConnectome database

The PlantConnectome is hosted on a Google Cloud server. The backend was implemented using the Python framework Flask and the Python packages networkx version 3.1, pickle version 3.11.4, json version 3.11.4, and regex *version 3.11.4*. We used JavaScript dependencies jQuery v3.6, Cytoscape.js v3.23, ChartJS v4.3, and FileSaver v2.0.5 to visualize the KnowledgeNetwork graphs.

API for PlantConnectome

PlantConnectome is also equipped with an Application Programming Interface to allow users to conduct search queries remotely. The API accepts GET requests and is implemented using the same set of packages described earlier. For each successful call to PlantConnectome's API, a JSON object is returned, containing the functional abbreviations, GO terms, other nodes, and text summaries associated with the search query. To perform searches using the API, users can add "/api/<search type>/<search query>" to the web address, where "<search type>" and "<search query>" are placeholders representing the type of search and user's query, respectively.

Results

Semantic Analysis of 101,341 Paper Abstracts

To retrieve articles that describe gene functions, we first investigated which journals contain the highest sources of experimentally-characterized genes. Our analysis revealed that Plant Cell, Plant Physiology, and Plant Journal are the top three, followed by journals not constrained to plant research, such as PNAS and the Journal of Biological Chemistry (Figure S1). All abstracts published after 2005 were then acquired, resulting in a total of 101,341, involving plant-specific journals and plant kingdom-specific manuscripts from generalist journals. A considerable number of abstracts came from both old journals, such as Plant Physiology (established in 1924), New Phytologist (1902), and the Journal of Experimental Biology (1950), as well as newer journals, such as Frontiers in Plant Science (2010) (Figure 1A).

We determined the surveyed journals' discussion of cellular compartments, organs, and biological functions to assess their considered research topics. Most journals did not show particular specificity for any topic, except Nucleic Acid Research, which focused on 'chromatin organization' (Figure 1B). The most common topics were leaves, roots, plant reproduction, and responses to the environment (Figure 1C), while the most frequent topic pairs were root-auxin and drought-proline (Figure 1D, Table S1).

To visualize the relationships among abstracts, we constructed a 2D neighbor embedding tSNE plot (Macosko et al., 2015) using the recommended 40 perplexity and 1,000 iterations, which resulted in a stable layout (Figure S2). The plots demonstrate clear groupings by organ, subcellular compartments, and biological processes (Figure 1E); the majority of entities studied also exhibit clear grouping (Figures S3, S4, and S5). However, we also observed certain overlaps: articles discussing roots (Figure 1E, left panel, orange points), for

example, overlap with membrane studies (Figure 1E, top right panel, red points), and RNA biosynthesis (Figure 1E, bottom right panel).

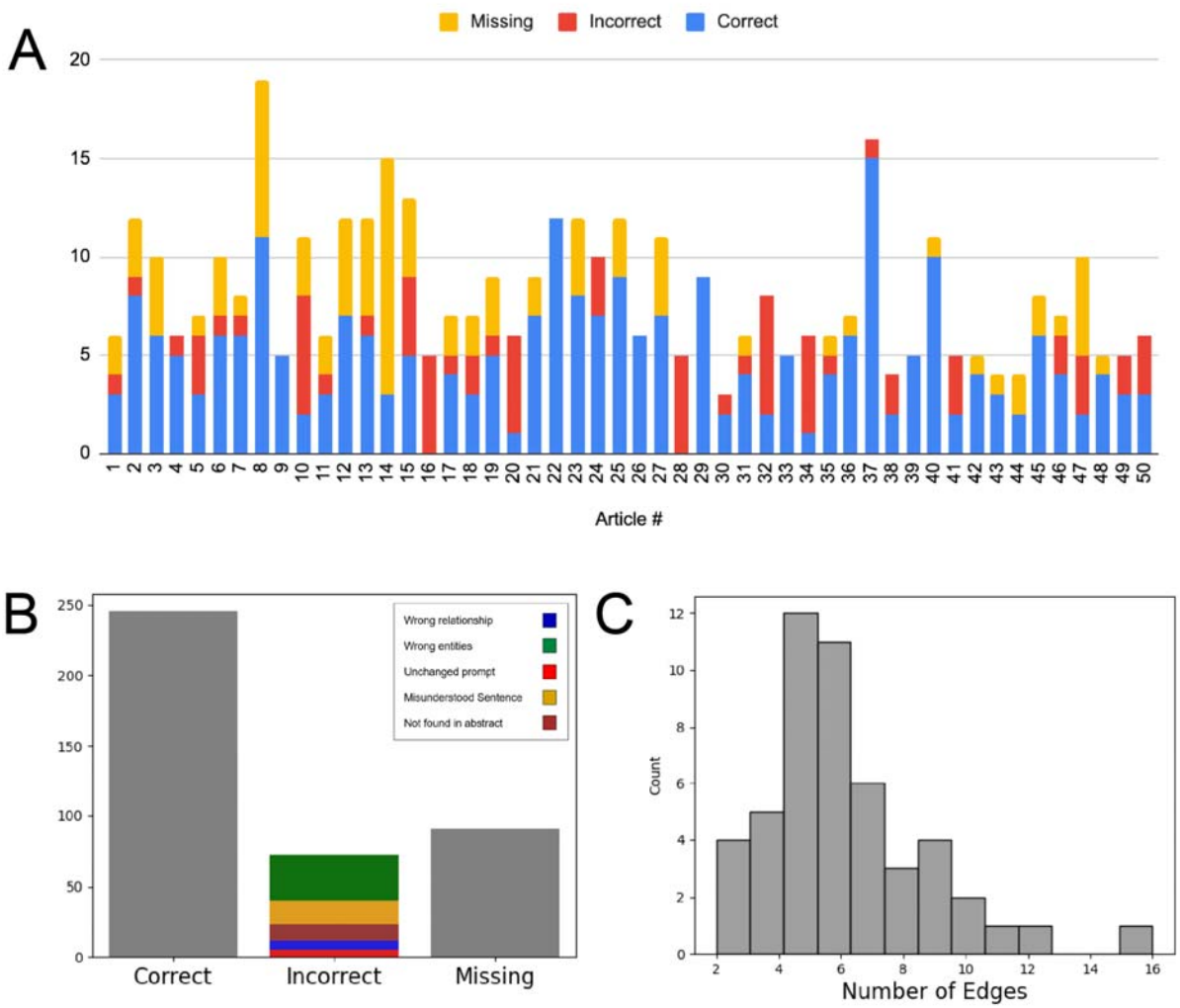

Figure 2. GPT analysis of abstracts. A) Distribution of correct, incorrect, and missing statements in fifty manually-inspected abstracts. B) Total number of correct, incorrect, and missing statements. C) Distribution of the number of edges across each of the randomly selected 50 abstracts.

Text-Mining Abstracts with GPT

We tasked OpenAI's GPT API with identifying functional relationships between pairs of entities (e.g., gene A interacts with gene B), proposing functional annotations of genes (gene A is involved in photosynthesis), and extracting any abbreviations (e.g., CESA is Cellulose Synthase A). To find the most effective prompt, we asked ChatGPT to propose a prompt for a given task. After several iterations, we arrived at the three prompts (Table 1), which were used to process all abstracts, yielding 387,777 relationships (Table S2), 112,128 function annotations (Table S3), and 73,591 abbreviations (Table S4).

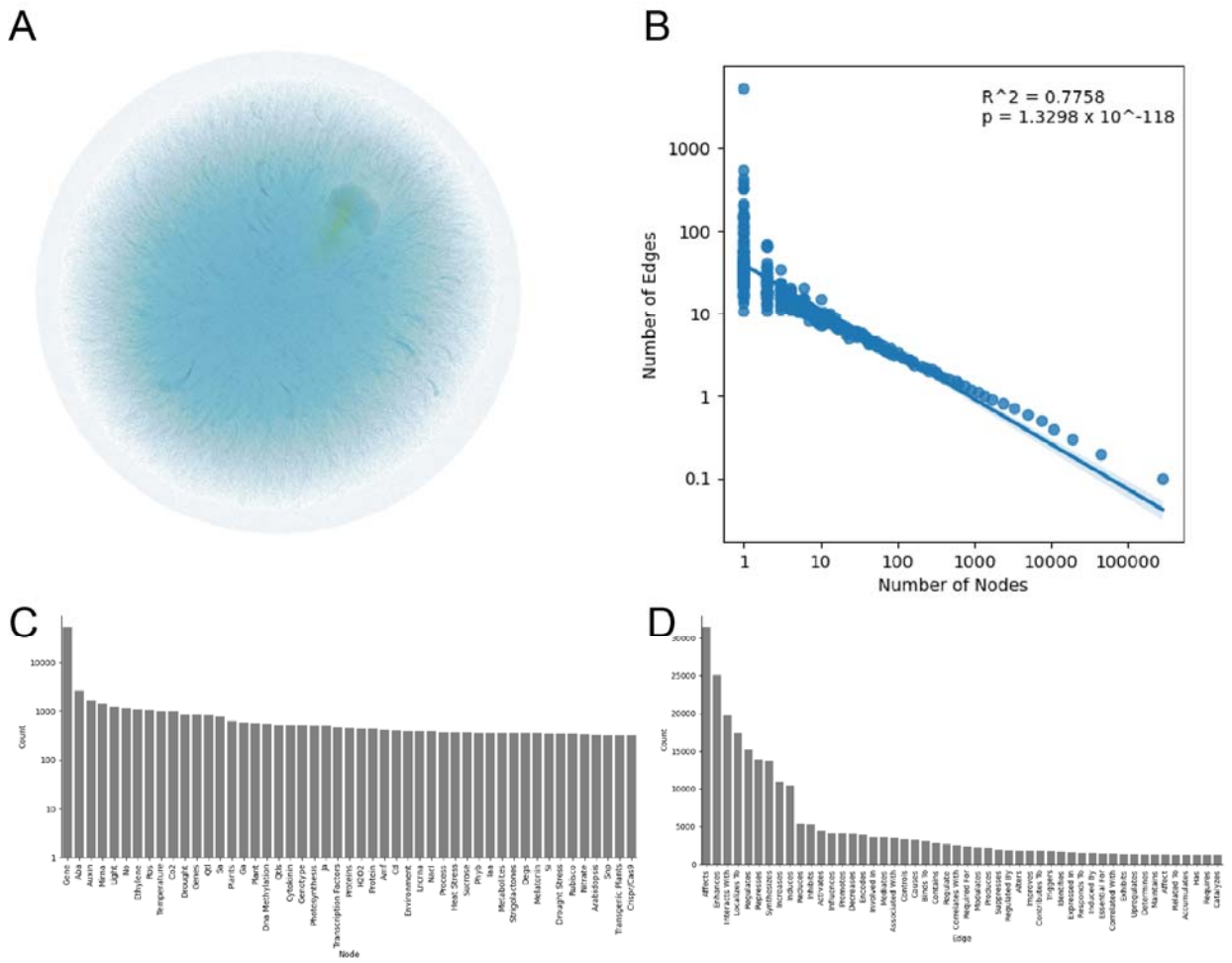

Figure 3: Properties of the Connectome network. A) Gephi visualization of the network. For the layout, we used the ForceAtlas 2 algorithm until convergence with a stronger gravity law and a scaling factor of 0.5. Light blue nodes and green nodes represent those with the fewest and highest degrees, respectively. B) Regression plot of the number of edges (y-axis) versus the number of nodes (x-axis) in the Connectome network. C) Top 50 most frequently-appearing nodes. The y-axis is log-transformed. D) Top 50 most frequently-appearing edges.

To quantitatively benchmark the accuracy of these results, we randomly selected 50 abstracts and manually evaluated the number of correct, incorrect, and missing relationships identified by GPT (Supplemental Data 1). Our analysis revealed that, overall, the majority of relationships were correct (Figure 2A, blue bars, Table S5). Still, a fraction of relationships remained undetected (orange bars) or wrong (red bars), inciting our defining of five additional categories within the inaccurate results: incorrect entities, incorrect relationship type, misunderstood sentence (i.e., both the involved entities and relationship type were misidentified), returned prompt (i.e., an output identical to the inputted prompt), and not found in the abstract. The predominating error type was ‘wrong entity’ (Figure 2B), suggesting GPT's limitations in correctly identifying entities discussed in the given text (Figure S6). Closer inspection of the associated abstract revealed that this error type occurs when it mentions no gene names, resulting in GPT's "hallucination" of an entity termed "gene" (Supplemental Data 1, Table S5,

abstracts 16 and 28). That we were able to extract around 4-6 relationships from each abstract (Figure 2C) indicates our overall ability to harvest information from the abstracts.

Construction of the Connectome Network

As GPT returned pairwise relationships between entities, we leveraged this to construct a network summarizing all entity-entity relationships. This network–the *PlantConnectome*–comprises regions of dense node clusters (Figure 3A). Certain networks, such as protein-protein interactions, display scale-free behavior, where most nodes have few connections and few nodes have many connections (Broido and Clauset, 2019). To investigate whether the Connectome is scale-free, we constructed a scatterplot of its log-transformed node frequency and node degrees, observing that it resembles a scale-free pattern (Figure 3B).

While GPT was instructed to focus on genes (Table 1), it still identified genes, hormones, metabolites, organs, and other biological entities (Figure 3C). Because many abstracts discussed gene function without explicitly stating the gene's name, "gene" was the Connectome's mode node (Figure 3C). The second, third, and fourth most frequently occurring nodes, however, were the plant hormones "aba," "auxin," and "ethylene," respectively. Importantly, GPT also recognized the types of relationships, where some of these most frequent edges were ‘affects’, ‘enhances’, and ’interacts with’ (Figure 3D).

Features of PlantConnectome

To provide access to the Connectome network, we constructed PlantConnectome (https://connectome.plant.tools/), which offers numerous methods of searching for genes, metabolites, organs, and other entities by terms, author names, and PubMed IDs, alongside a catalog page (accessible under the "entities" tab) listing all entities in the database. An entire information page is also provided for each entity in the connectome, containing its GPT-generated abbreviations and GO term predictions in addition to appropriate links. To make the PlantConnectome easier to use, we manually identified synonymous edges (e.g., ENCODE, CODES FOR, ENCODES FOR, CODE FOR become ENCODE, Table S6).

To detail PlantConnectome's search result page, we performed a standard query with the gene "CESA" (cellulose synthase A, https://connectome.plant.tools/normal/CESA), which is involved in the biosynthesis of primary and secondary cell walls of plants (Lampugnani et al., 2019). Following a statement of the total number of contributing publications is an "abbreviations" section, listing all descriptions of CESA abbreviations (Figure 4). The subsequent KnowledgeNetwork is a visual depiction of the various relationships the search query shares with other entities in the database. Upon selection of a given node, the user is provided a tooltip displaying the node’s abbreviations, functional annotations, and a set of options enabling removal of the node, isolation of the node’s neighborhood, and visiting the node's corresponding entity page. Users may, thus, customize the network by erasing nodes or entire clusters of choice and/or filtering out specific relationship types (e.g., "activates," "binds," "encodes for," etc.), enabled by the "Layout Option" button; to facilitate this process, a search box is also provided above the network. As an excessively large network is associated with certain entities (e.g., "heat" has 970 papers https://connectome.plant.tools/normal/heat), the networks are limited to 500 nodes; nonetheless, the user can still download the full version as a tab-delimited file (visualizable in Cytoscape, for example).

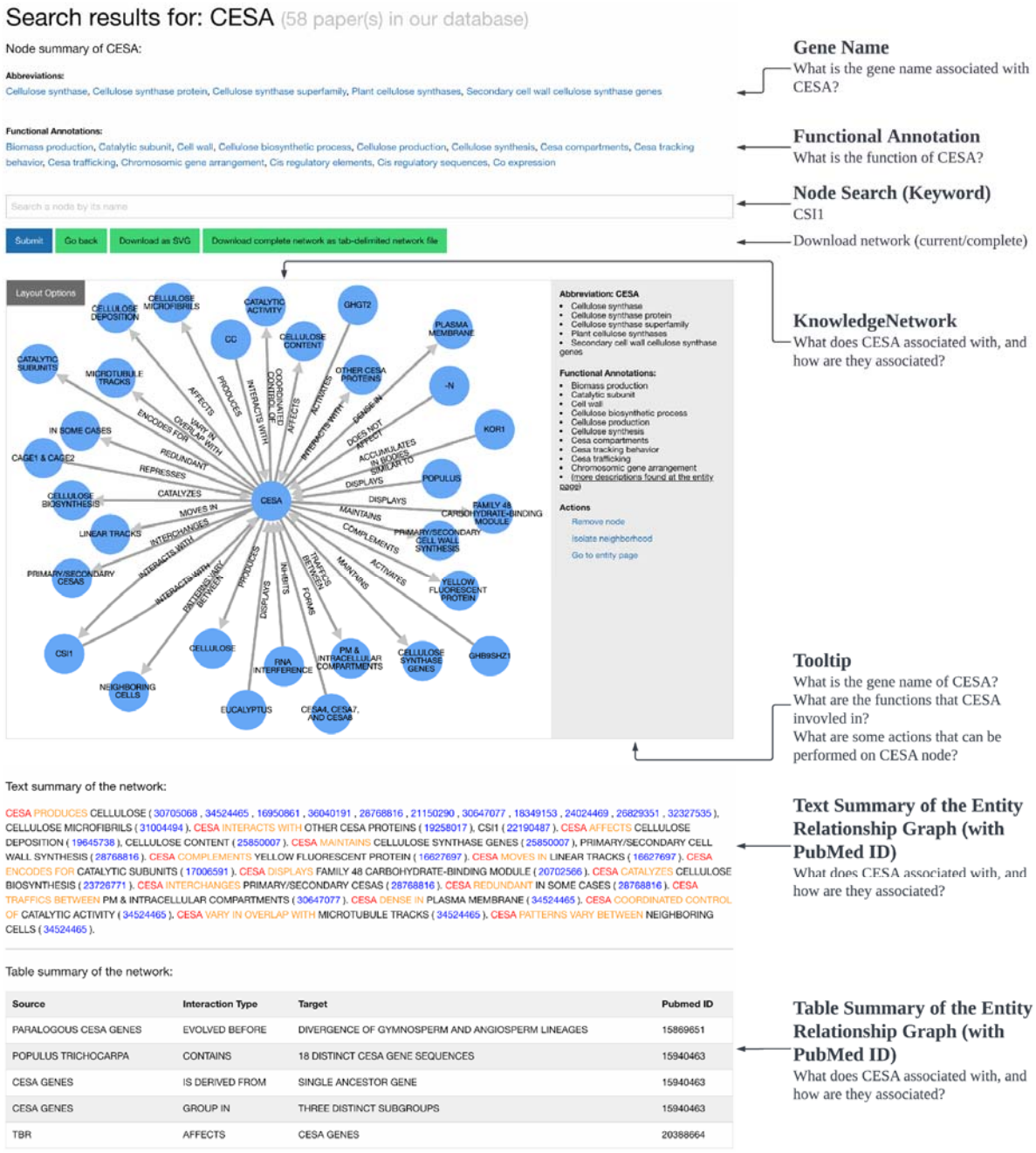

Figure 4. Outline of the ConnectomeDatabase gene search page contents. After inputting our query entity (CESA), the result page returns information specific to the searched term(s), including gene names, functions, entity relationship graph, as well as a text and table summary. Users may also save the network as a vector graphic (i.e., SVG) or tab-delimited file (i.e., TSV).

Below the KnowledgeNetwork is its text summary, arranged in order of decreasing node degree (i.e., the number of other entities to which a given node is connected). Each paragraph contains the entities, corresponding association types, and PubMed IDs of the publications from which the relationships were mined. Clicking on a given PubMed ID will prompt a popup

containing the corresponding abstract and published article link, streamlining the manual confirmation of relationships (as warranted by GPT's potential errors, Figure 2). Selecting edges/nodes will condense this text summary to the said selection, providing users with an expeditious means of gathering information on relationships of interest. The network can also be found in the table representation at the lowermost section of the results page.

Finally, PlantConnectome enables users to perform searches through an API, which returns a JSON object containing relevant network and functional information, extending its functionality to bioinformaticians who desire programmatic access to our database. As an example, a normal search on the CESA gene may be performed by accessing the URL "https://connectome.plant.tools/api/normal/cesa".

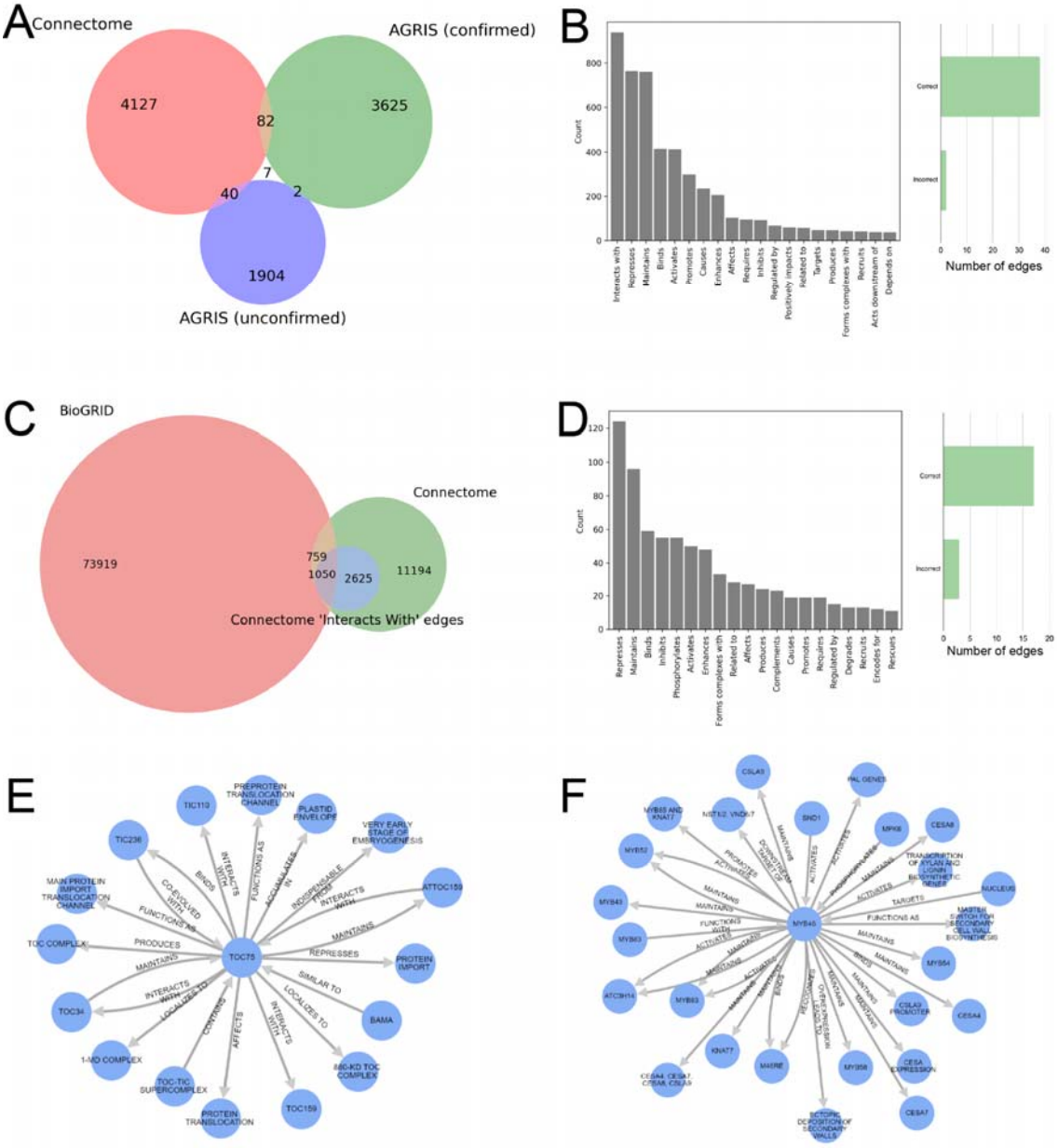

Figure 5. Comparative analysis of Connectome's gene regulatory networks with AGRIS and protein-protein interaction networks with BioGRID. A) Venn diagram showing the

intersection of Connectome's and AGRIS' gene regulatory networks. While AGRIS (confirmed) refers to gene regulatory networks that have been validated experimentally, AGRIS (unconfirmed) refers to those that are suggested by large-scale studies without experimental evidence of gene regulation by the respective TF. Networks formed by TFs that are not found in both Connectome and AGRIS were excluded to ensure fair comparison. B) Top 20 types of Connectome's edges that do not overlap with AGRIS. C) Venn diagram showing the intersection of Connectome's and BioGRID's protein-protein interaction networks. D) Top 20 types of overlapping edges between Connectome and BioGRID, excluding the "interacts with" edges. E) Knowledge Network of MYB46. F) Knowledge Network of TOC75.

Evaluation of the coverage and accuracy of the PlantConnectome

Our main motivation in this study was to expand the amount of the gold standard data capturing experimentally-verified gene functions. We, thus, investigated the overlap of relationships detected by GPT with data provided in the public repositories.

To compare the coverage and the accuracy of gene regulatory networks (GRNs), we obtained the *Arabidopsis thaliana* gene regulatory network from AGRIS (https://agris-knowledgebase.org/downloads.html, updated March 2019) (Yilmaz et al., 2011). Next, we identified 4,249 transcription target-gene edges in the Connectome and calculated the overlap between AGRIS and Connectome. A very minor overlap between them was observed (Figure 5A), indicating the high dissimilarity between the two networks and that Connectome plays a complementary role to AGRIS.

To understand the types of edges associated with transcription factors in the Connectome, we investigated these edges, finding that "interacts with," "represses," "maintains," "binds," and "activates" were the top five, which demonstrates the Connectome's ability to identify the various functions of transcription factors (Figure 5B). A total of 40 edges were also randomly sampled to validate the edge-detection accuracy of the Connectome, where 95% of them were correct (Table S7), corroborating that the Connectome is an accurate, complementary companion to AGRIS (Figure 5C).

Furthermore, we compared the protein-protein (PPI) network from BioGRID to the Connectome's network. While there was a greater overlap between the two networks (1,050 out of 3,675 of 'interacts with' edges), Connectome was still able to identify 2,625 interaction edges that were not found in BioGRID (Figure 5C). Additionally, 759 edges that overlapped with BioGRID but were not of the "interacts with" type was detected by the Connectome, upon examination of which we discerned that more niche interactions types, such as "phosphorylates" and "inhibits," were, too, captured by the Connectome. After manually inspecting 20 randomly-sampled edges, we determined an edge accuracy rate of 85% (Figure 5D, Table S8), indicating the Connectome's valuable companionship and alternative role to not only AGRIS but also BioGRID. A full list of all entity pairs that are assigned Arabidopsis gene identifies are available (Table S9).

Examples of how to use PlantConnectome

In building PlantConnectome, we have distilled the knowledge from the majority of research abstracts, providing a versatile tool to the plant community covering diverse topics. We now

demonstrate how the Connectome can be used to study gene regulatory networks, protein complexes, cell types, and stress resilience.

Example 1: Secondary Cell Wall Master Regulator

In this example, we selected the secondary cell wall biosynthesis regulator, *MYB46* (https://connectome.plant.tools/normal/myb46). A literature search on the gene regulatory network underlying secondary cell wall formation revealed that Connectome has a GRN rivaling those in a review regarding this topic (Figure 5E) (Xiao et al., 2021). While the KnowledgeNetwork viewer did not contain several other transcription factors that work alongside *MYB46* (*MYB42, MYB58, MYB103, MYB7*, and *MYB32*), we nonetheless observed transcription factors *SND1, MYB43, MYB52, MYB46, MYB85, KNAT7*, and *MYB63* (Figure 5A).

Since the KnowledgeNetwork is summarized in a human- and machine-readable format, we investigated whether ChatGPT is able to generate publication-ready passages. To this end, the network summary (Figure 4, below the network) was fed into ChatGPT4 with the following prompt: "*I will give you several statements about gene function, where the numbers in "()" contain PubMed IDs. Rewrite the text so that it reads like a scientific article, and group logically the different statements. Make sure to keep all PubMed IDs.*" The outcome is, in our opinion, on par with some review articles:

> *Introduction*:
> The *MYB46* transcription factor is essential in the regulation of secondary cell wall biosynthesis in plants. It has been characterized as a master switch that integrates with cell wall remodeling and defense-related signaling pathways (PubMed ID: 33826618, 21282403). This article presents an overview of the role of *MYB46* in secondary cell wall biosynthesis, as well as its interaction with other genes and regulatory elements.
>
> *Expression and Function of MYB46:*
> *MYB46* is predominantly expressed in fibers and vessels (PubMed ID: 17890373). The primary function of *MYB46* involves the maintenance of secondary wall biosynthesis through the induction of several genes including *EP5C* (PubMed ID: 19838072, 21282403), *MYB43, MYB52, MYB54, MYB58, MYB63, KNAT7, ATC3H14*, and genes involved in cellulose synthesis such as *CESA* and *CSLA9* (PubMed ID: 22197883, 22271306, 23726771, 24243147, 24879533, 26011122). Additionally, *MYB46* promotes the transcription of xylan and lignin biosynthetic genes, *PAL* genes, and affects xylose and lignin content, as well as crystalline cellulose content (PubMed ID: 24879533, 26011122, 32706429).
>
> *MYB46 Activation and Binding:*
> *MYB46* has the ability to activate several genes including *ATC3H14, MYB52, MYB63*, and secondary wall biosynthetic genes (PubMed ID: 19674407, 22197883, 24879533). It achieves this by binding to various regulatory elements such as *EP5C* gene, *SMRE*, the native promoter of *CSLA9*, and promoters of lignin and xylan biosynthesis genes (PubMed ID: 21282403, 22197883, 23726771, 24243147, 24879533, 26011122).

Interestingly, *MYB46* recognizes a specific element known as M46RE, which also has been shown to inhibit MYB46 binding (PubMed ID: 22271306, 23726771).

Interactions with Other Factors:
MYB46 is transcriptionally reprogrammed by *SND1*, which activates *MYB46* and binds to its promoter (PubMed ID: 17890373). Moreover, *MYB46* operates in conjunction with *MYB83*, activating similar direct targets (PubMed ID: 19808805, 22197883). *MYB46* is also subject to repression by *AN* and *miR395c* (PubMed ID: 32706429, 35755696). Additionally, *MYB46* is targeted to the nucleus (PubMed ID: 17890373). *MPK6* is another regulatory protein that interacts with *MYB46* by phosphorylating it, which leads to the degradation of *MYB46* and helps to maintain the in planta secondary wall-forming function of *MYB46* (PubMed ID: 33826618).

Impact of MYB46 Regulation:
When overexpressed, *MYB46* leads to ectopic deposition of secondary walls (PubMed ID: 17890373). However, it has also been shown to repress secondary wall thickening (PubMed ID: 17890373). In *MYB46* mutant plants, selective transcriptional reprogramming of cell wall proteins and extracellular remodeling enzymes occurs (PubMed ID: 21617373). Furthermore, *MYB46* is identified as a downstream target of *NST1/2* and *VND6/7*, which play crucial roles in secondary wall biosynthesis (PubMed ID: 19122102).

Since ChatGPT was asked to use the existing Pubmed IDs, it does not hallucinate non-existing references.

Example 2: Chloroplast Protein Translocation and Channel Member TOC75
Translocase complexes on the outer and inner envelope membranes (TOC and TIC, respectively) are used to import proteins into the chloroplast (Stengel et al., 2009). We conducted a search on "*TOC75*" (https://connectome.plant.tools/normal/TOC75), where comparison of the output to a review revealed known interactions with other translocase components, such as *TOC34*, *TOC159*, and *TIC236* (Figure 5E) (Richardson and Schnell, 2020). The associated edges also provide additional functional information: its role (i.e., a pre-protein translocation channel), supercomplex membership, subcellular compartments, and mutant phenotypes. Only one known interactor - *TIC22* - was not found in the network. Furthermore, while Connectome states that *TOC75* is implied to interact with *TIC110* in the network, these two proteins do not interact directly (Richardson and Schnell, 2020). Closer inspection of the sentence in the corresponding abstract, which states, "Antibody-shift assays showed that the 1-MD complex is a TOC-TIC supercomplex containing at least *Toc75*, *Toc159*, *Toc34* and *Tic110*", implies that GPT likely inferred that all proteins in the supercomplex were interacting. Additionally, the identified relationship "*TOC75* represses protein import" is incorrect, as the downregulation of *TOC75* represses protein import rather than *TOC75* activity. Regardless, because PlantConnectome significantly simplifies the process of scrutinizing the abstracts underpinning all edges, such inconsistencies are easy to spot.

Example 3: Root Hair Development

Root hair development can be described in several stages and is influenced by a magnitude of external and internal factors (Shibata and Sugimoto, 2019). The word "hair" was searched on PlantConnectome (https://connectome.plant.tools/normal/hair), the result of which (based on 397 papers) revealed a high number of connected nodes such as "root hair development", "root hair initiation", and "root hair elongation". To identify entities that are important for root hair maintenance, we selected the "maintains" and "affects" edges, reducing the network's complexity (Figure S7). This subsequent network revealed an extensive amount of internal and external factors such as calcium (30153078), reactive oxygen species (16720604), soil acidity (21062319), growth media (21062319), phytohormones ethylene (30153078, 16531464), auxin (35401627, 27799284), and jasmonic acid (35401627). Genes such as *KOJAK/ATCSLD3* (17259288), *ETC1* (23432399), *NPC4* (23432399), and *SQD2* (23432399) were also revealed in the network. These results demonstrate that PlantConnectome is able to integrate knowledge across different types of entities.

Example 4: Thermotolerance

Thermotolerance is a widely studied phenomenon that enables plants to withstand high temperatures (Ali et al., 2020). We navigated to the following search results page https://connectome.plant.tools/substring/thermotolerance, which revealed 397 publications on the topic. To identify entities that promote thermotolerance, we selected the network's "enhances" edges. Both a range of molecules, such as isoprenes (17468218), DGDG (Digalactosyl diacylglycerol, 17080965), abscisic acid (15923322), and nitric oxide (18326829), and a host of genetic factors, such as *HSFA2* (30187931), *HSA32* (24520156), *HOT1* (10760305), *HSP101* (24520156) (Figure S7), were illuminated.

The network can also be used to reveal entities that inhibit thermotolerance. Namely, by selecting "represses" and "inhibits" edges, we determined that MGDG (Monogalactosyldiacylglycerol, 17080965) is an inhibitor of thermotolerance. A closer inspection of the abstract suggested that a decreased ratio of DGDG to MGDG is responsible for lowering thermotolerance; thus, while the network's implications are not strictly correct, MGDG is still implied to play a role in thermotolerance. Moreover, the network revealed several genes, such as *CLPC1* (33326777) and *SLMAPk3* (31412779) that repress thermotolerance, and while the genes *HSA32* (16500991) and *APX2* (33711164) were implied to be thermotolerance repressors, manual review of the abstract exposed that GPT missed that it was these genes' mutants that were discussed and that the two genes actually enhanced thermotolerance. To conclude, the dynamic selection option of edge types in the network enables scrutinizing different relationship types between the entities found in PlantConnectome.

Discussion

We have illustrated GPT's text mining capacities in the context of scientific literature, processing over 100,000 research abstracts at a moderate cost (1,500 USD) within two weeks and harvesting invaluable functional information therein. GPT was capable of extracting key entities and relationships from research paper abstracts with high accuracy and few prompts (Table 1, Figure 2). The amount of functional information excavated from the abstracts vastly increased the amount of machine-readable data, as demonstrated by our gene regulatory networks that

doubled the quantity of available data. Moreover, PlantConnectome overcomes the limitations of typical databases that employ only one data type, as it draws upon numerous sources of data in its establishing of gene functions, organ development, gene regulatory networks, protein-protein interactions, and other phenomena, all in a user-friendly manner.

Our evaluation has shown that PlantConnectome is not only comprehensive and accurate but complementary to existing databases (Figure 5). The comparison of PlantConnectome's gene regulatory networks against AGRIS and its protein-protein interaction networks against BioGRIDdemonstrate that PlantConnectome's retrieved networks do not largely overlap with these reference databases. Rather, the GPT-extracted networks complement them, bearing witness to the effectiveness of our text mining approach in utilizing the vast amount of literature that has not been captured by manual curation.

However, GPT's outputs are not entirely accurate and still warrant manual verification, of which our own has shown that the OpenAI model has a tendency to misidentify entities and relationships or to not detect them at all (Figure 2) which is perhaps attributable to each abstract's varying language and content. The correction of errors may be carried out by fine-tuning the models with manually curated examples containing the expected output (as, for instance, that found in Supplemental Data 2). Moreover, while we asked GPT to annotate genes with GO terms, GPT, in many cases, formulated new terms; for example, 'Preprotein translocation channel,' assigned to TOC75, is a valid but non-existing term (https://connectome.plant.tools/normal/TOC75). A possible means to address this could involve using Natural Language processing methods that match GPT annotations with valid GO terms (Wang et al., 2020).

In conclusion, PlantConnectome is an innovative tool, combining the power of a state-of-the-art language model with the comprehensive information embedded in a massive collection of research articles. The tool offers an efficient and diversified way to retrieve information for genes, metabolites, tissues, organs, and other biological components. The potential applications of PlantConnectome are wide-ranging and extend beyond those we have highlighted in this article. Furthermore, since we only analyzed article abstracts, we anticipate that analysis of complete articles will return even more information. We anticipate that PlantConnectome will become a valuable resource for the plant science community to facilitate various research activities, from a preliminary investigation of gene functions to an in-depth study of a particular biological process.

Tables

Table 1. An example of an abstract, prompts and outputs from GPT. The three prompts ask GPT to identify relationships between genes, identify Gene Ontology terms and find abbreviations, respectively.

Abstract	In plants, regulation of cellulose synthesis is fundamental for morphogenesis and plant growth. Cellulose is synthesized at the plasma membrane, and the orientation of synthesis is guided by cortical microtubules; however, the guiding mechanism is currently unknown. We show that the conditional root elongation pom2 mutants are impaired in cell elongation, fertility, and microtubule-related functions. Map-based cloning of the POM-POM2 locus revealed that it is allelic to CELLULOSE SYNTHASE INTERACTING1 (CSI1). Fluorescently tagged POM2/CSI1s associated with both plasma membrane-located cellulose synthases (CESAs) and post-Golgi CESA-containing compartments. Interestingly, while CESA insertions coincided with cortical microtubules in the pom2/csi1 mutants, the microtubule-defined movement of the CESAs was significantly reduced in the mutant. We propose that POM2/CSI1 provides a scaffold between the CESAs and cortical microtubules that guide cellulose synthesis.		
Prompt	Write a very short summary about the functions of genes in this abstract. The summary must show pair-wise relationships, for example: gene: !affects! Process gene: !localizes to! X gene: !interacts with! Y gene: !enhances! Z gene: !represses! U gene: !synthesizes! I Please provide only one statement per line, and ensure that each line contains exactly two actors. If a relationship involves more than two actors, please break it down into multiple separate lines. <ABSTRACT> VERY SHORT, CONCISE SUMMARY CONTAINING ALL INFORMATION WITH TWO ACTORS PER LINE:	Your job is to identify GO terms of genes from scientific abstracts. For example, given an abstract: 'LUCKY localizes to plasma membrane, where it interacts with the extracellular protein HAPPY' You should return: LUCKY: plasma membrane, protein binding HAPPY: extracellular region, protein binding For each gene in this abstract, annotate it with fitting GO Terms. <ABSTRACT>	Return abbreviations and their corresponding definitions from scientific abstracts. The input to the function should be a string representing the abstract, and the output should be a dictionary where the keys are the abbreviations and the values are their definitions. For example: Input: "Cellulose synthase complexes (CSCs) at the plasma membrane (PM) are aligned with cortical microtubules (MTs) and direct the biosynthesis of cellulose. We identified a unique small molecule, CESA TRAFFICKING INHIBITOR (CESTRIN), which reduces cellulose content and alters the anisotropic growth of Arabidopsis (Arabidopsis thaliana) hypocotyls." Output: {'CSCs': 'Cellulose synthase complexes', 'PM': 'plasma membrane', 'MTs': 'cortical microtubules', 'CESTRIN': 'CESA TRAFFICKING INHIBITOR'} <ABSTRACT>
Output	Cellulose: !synthesized at! Plasma Membrane Cortical Microtubules: !guide! Cellulose Synthesis POM2/CSI1: !associates with! CESAs POM2/CSI1: !provides scaffold between! CESAs and Cortical Microtubules	CELLULOSE SYNTHASE INTERACTING1 (CSI1): plasma membrane, protein binding, microtubule binding CESAs: plasma membrane, cellulose biosynthetic process, microtubule binding	{'CESAs': 'plasma membrane-located cellulose synthases', 'CSI1': 'CELLULOSE SYNTHASE INTERACTING1', 'POM2': 'conditional root elongation pom2 mutants'}

Supplemental Figures

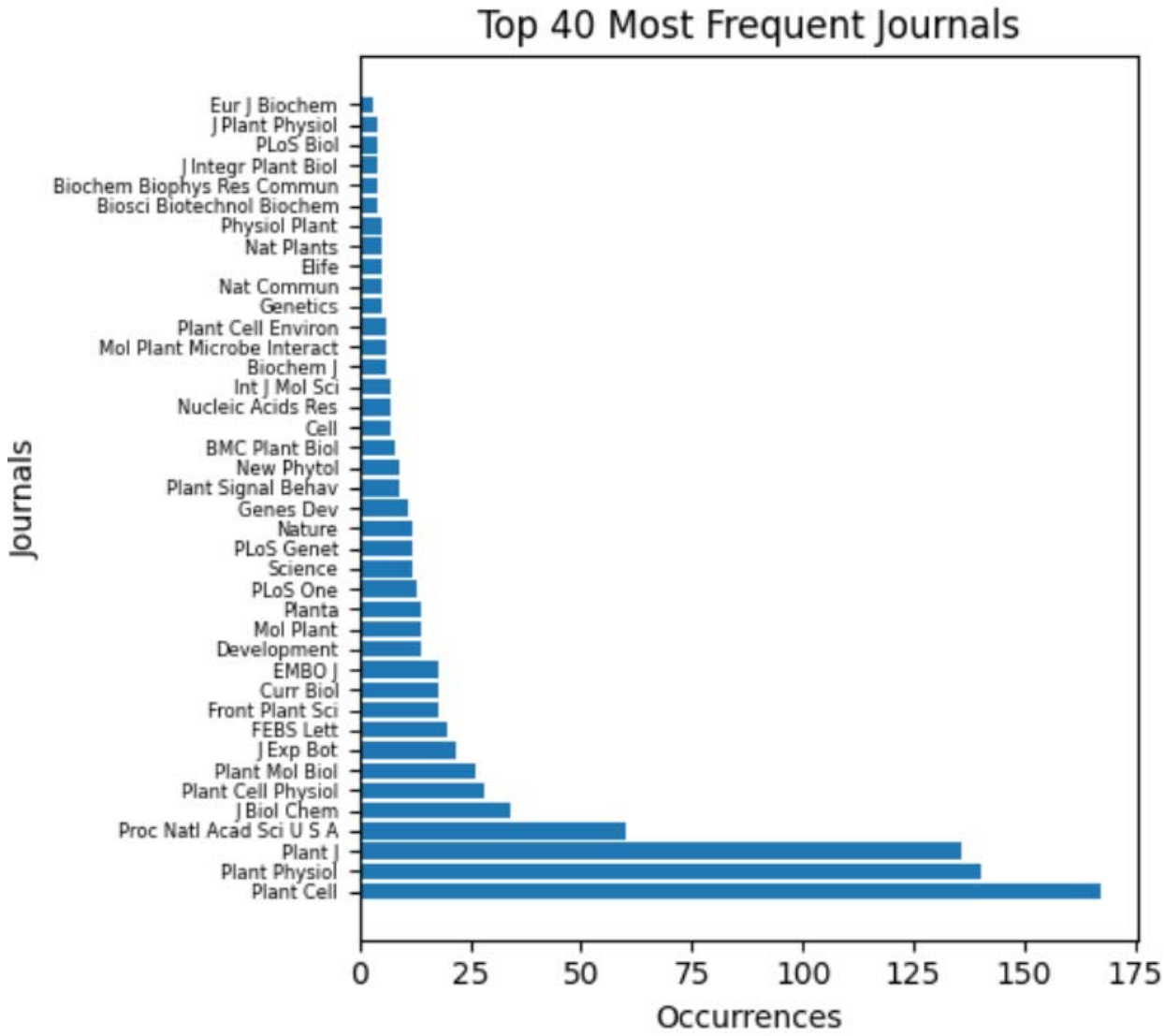

Figure S1. The number of articles that contain experimentally described genes. Experimentally described genes were downloaded from https://arabidopsis.org/download_files/GO_and_PO_Annotations/Gene_Ontology_Annotations/ATH_GO_GOSLIM.txt.gz

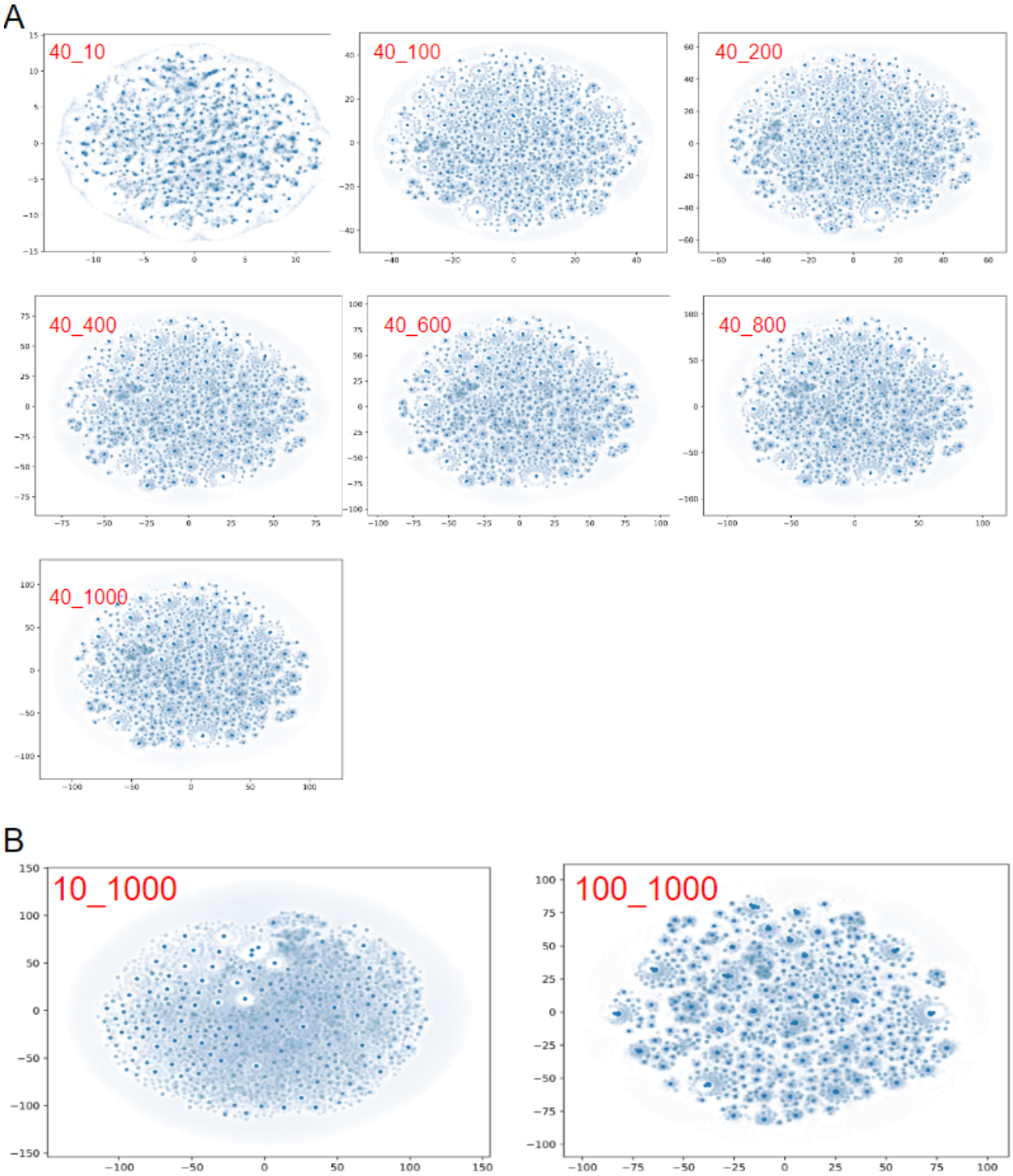

Figure S2. tSNE analysis of the abstracts at the different perplexity and iteration values. A) The evolution of the plot at a perplexity of 40 and different ranges or iterations. The plot at 1000 iterations resembles the plot at 800 iterations, indicating a plateau. B) The plot at a low perplexity (10) and maximum perplexity (100).

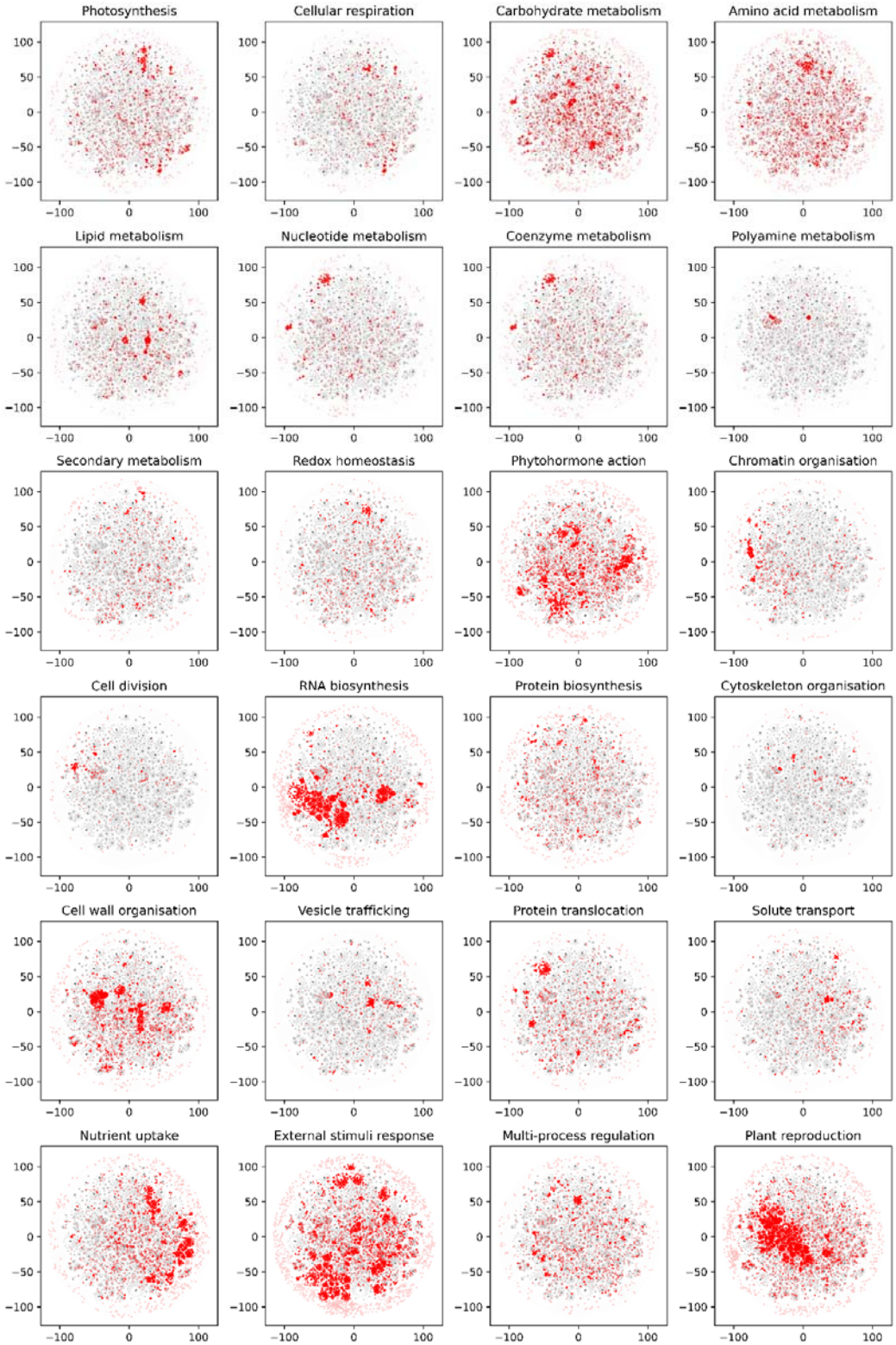

Figure S3. tSNE analysis of the abstracts of the different biological processes, as defined by MapMan. A red point indicates an abstract that contains a keyword (e.g., pollen is a keyword for the plant reproduction), while gray point indicates an absence of the keyword match.

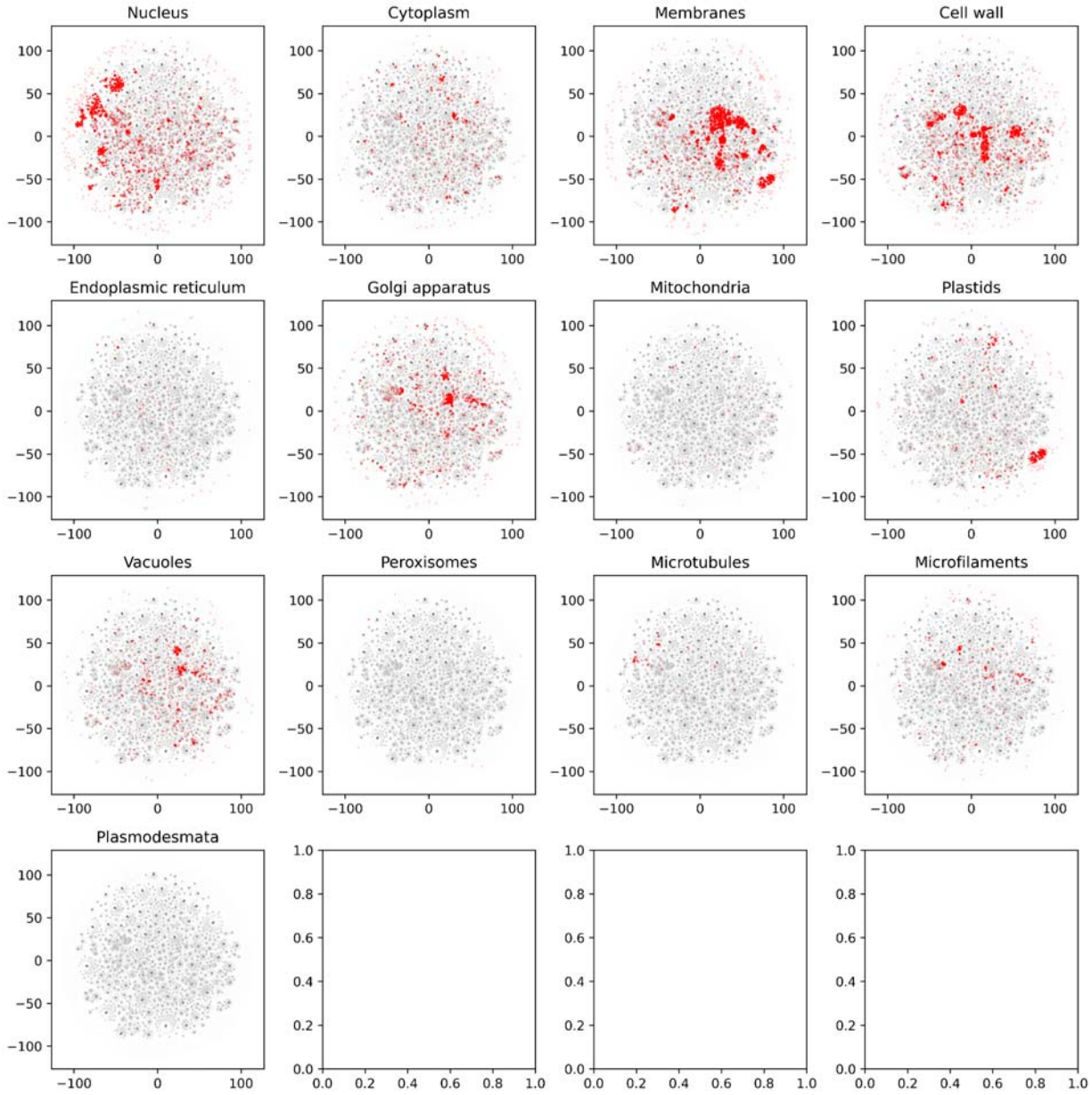

Figure S4. tSNE analysis of the abstracts of the different cellular compartments.

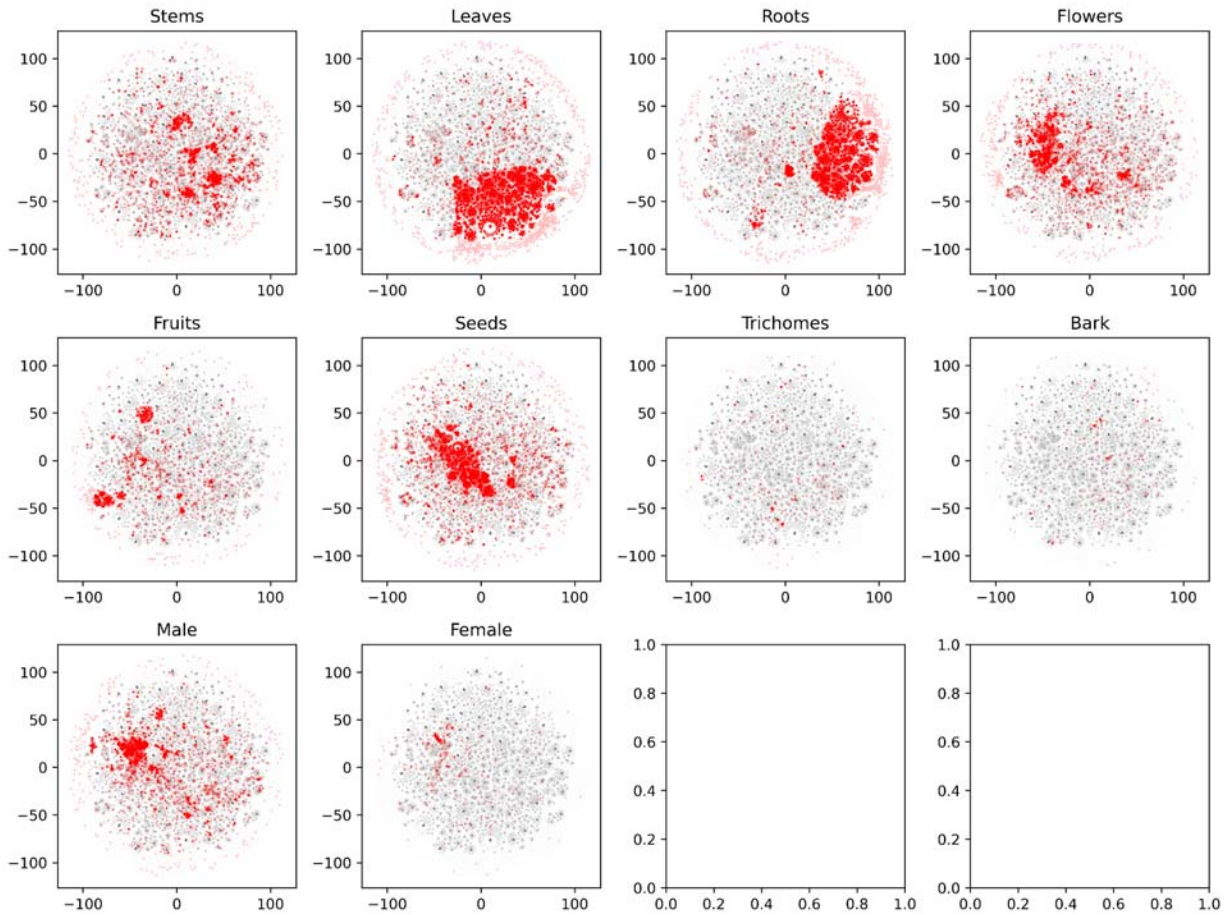

Figure S5. tSNE analysis of the abstracts of the different major organs and cell types.

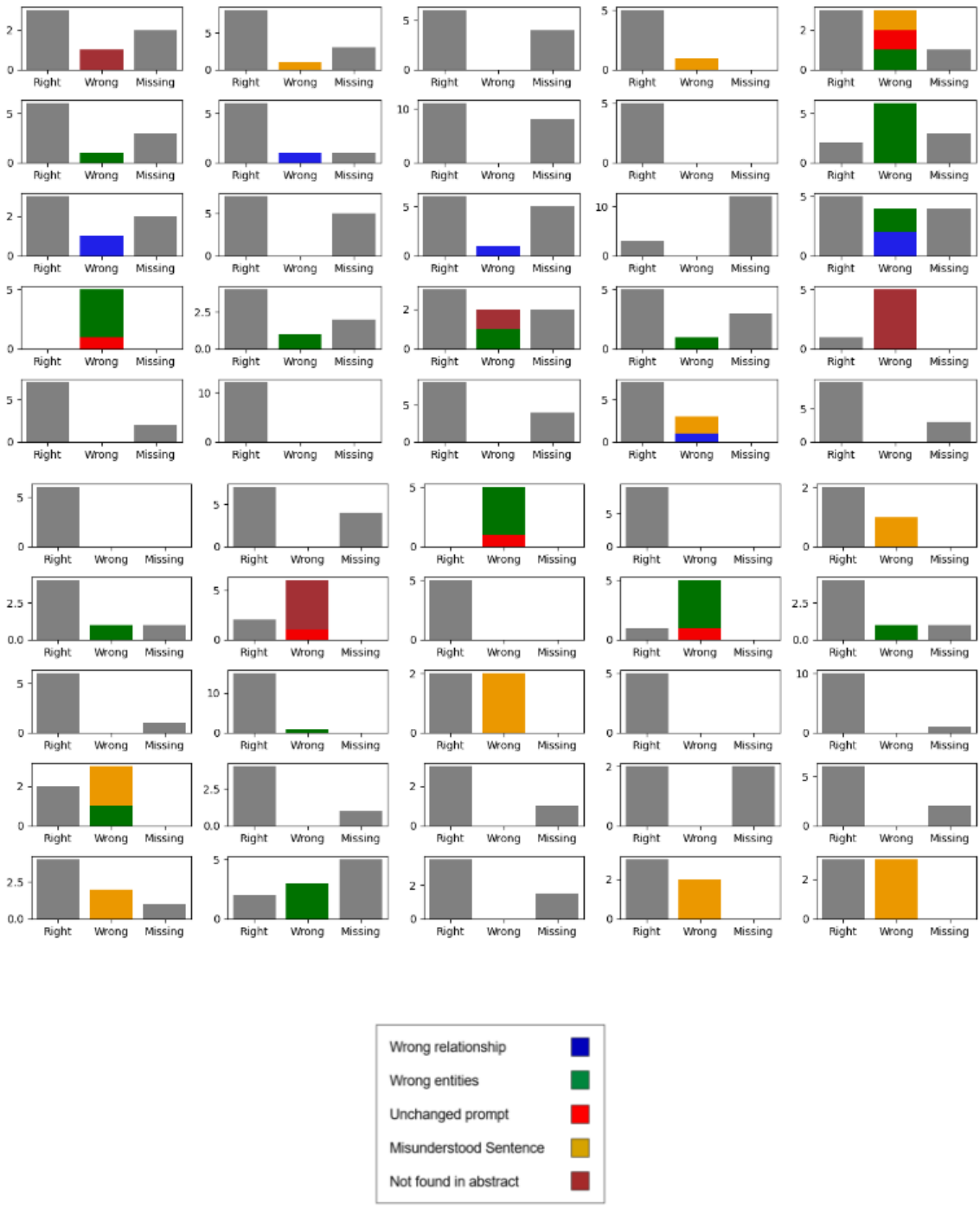

Figure S6. Manual curation of 50 abstracts.

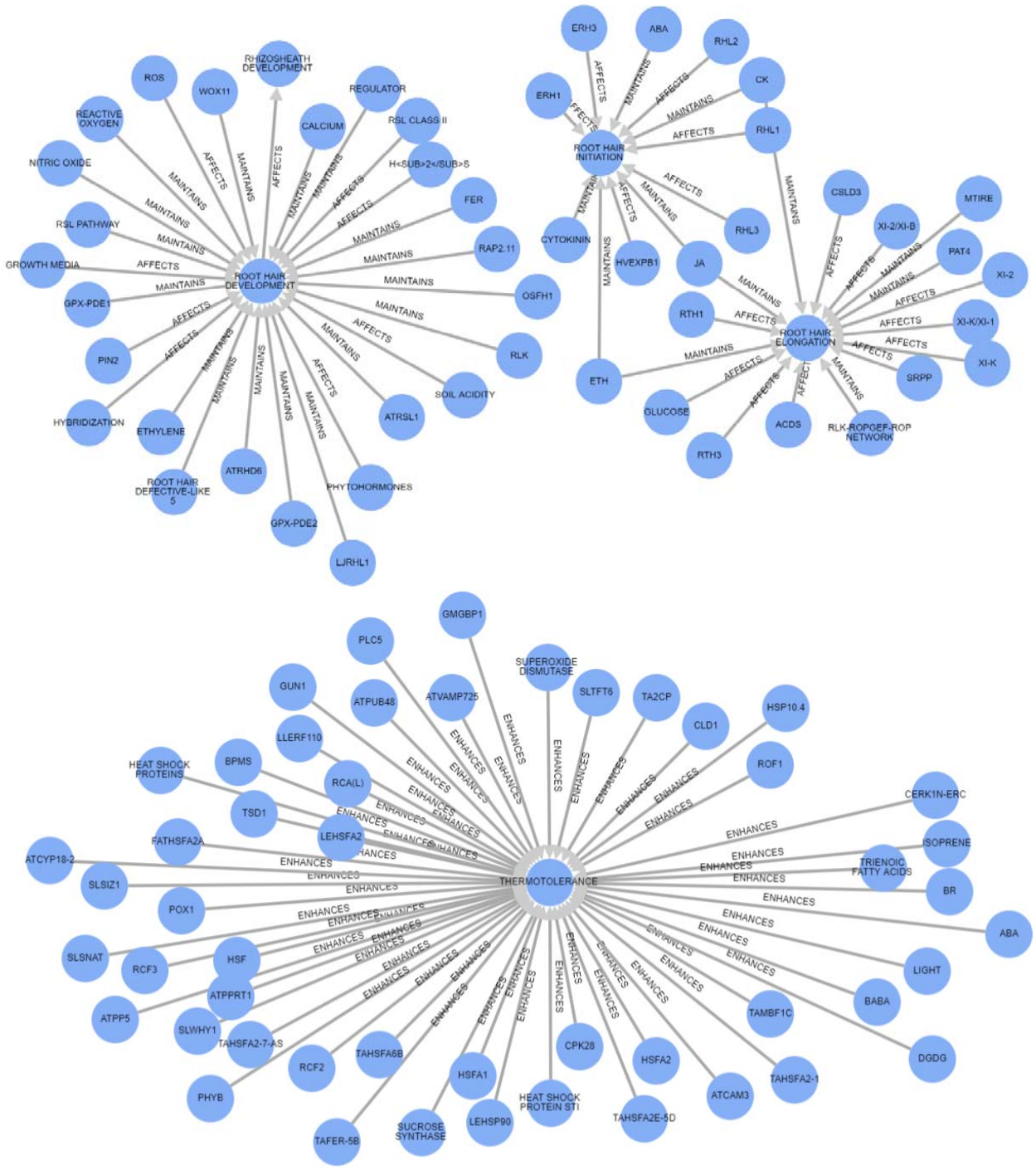

Figure S7. KnowledgeNetwork view of 'hair' (top) and 'thermotolerance' (bottom) networks.

Supplemental Tables

Table S1. Jaccard index between keywords found in the article abstracts. The pairs are sorted by the decreasing JI values.

Table S2. All entity relationships identified by GPT. The two entities are in columns A and C, while the edge type is specified in column B.

Table S3. Entity annotations identified by GPT. Column A contains the entity, while column B contains the outcome of asking GPT for GO terms. Each 'GO term' is separated by comma.

Table S4. Abbreviations identified by GPT. Column A and B contain the key and value of an abbreviation, respectively. Abbreviations identified from multiple abstracts are separated by comma.

Table S5. Accuracy summary of GPT inferences using 50 abstracts. The columns indicate the abstract ID, and the correct (B), incorrect (C) and missing (D) statements.

Table S6. Edge alias table. Column A shows the representative edge of the other edges (column B).

Table S7. Comparison of the GRN inferred by GPT and AGRIS. Columns I and J indicate whether an edge is found in AGRIS and if the edge is correct upon manual inspection, respectively.

Table S8. Comparison of the PPI inferred by GPT and BioGRID. Column J indicate whether an edge is correct upon manual inspection.

Table S9. The list of all entities that could be assigned AGI codes.

References

Ali, S., Rizwan, M., Arif, M.S., Ahmad, R., Hasanuzzaman, M., Ali, B., and Hussain, A. (2020). Approaches in Enhancing Thermotolerance in Plants: An Updated Review. J. Plant Growth Regul. **39**: 456–480.

Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., and Obayashi, T. (2016). ATTED-II in 2016: A plant coexpression database towards lineage-specific coexpression. Plant Cell Physiol. **57**: e5.

Broido, A.D. and Clauset, A. (2019). Scale-free networks are rare. Nat. Commun.

Brown, D.M., Zeef, L.A.H., Ellis, J., Goodacre, R., and Turner, S.R. (2005). Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell **17**: 2281–2295.

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G.D., and Morris, Q. (2018). GeneMANIA update 2018. Nucleic Acids Res. **46**: W60–W64.

Guo, J., Xu, W., Hu, Y., Huang, J., Zhao, Y., Zhang, L., Huang, C.-H., and Ma, H. (2020). Phylotranscriptomics in Cucurbitaceae Reveal Multiple Whole-Genome Duplications and Key Morphological and Molecular Innovations. Mol. Plant **13**: 1117–1133.

Julca, I. et al. (2021). Comparative transcriptomic analysis reveals conserved programmes underpinning organogenesis and reproduction in land plants. Nat. Plants **7**: 1143–1159.

Lampugnani, E.R., Flores-Sandoval, E., Tan, Q.W., Mutwil, M., Bowman, J.L., and Persson, S. (2019). Cellulose Synthesis – Central Components and Their Evolutionary Relationships. Trends Plant Sci. **24**: 402–412.

Lim, P.K., Zheng, X., Goh, J.C., and Mutwil, M. (2022). Exploiting plant transcriptomic databases: Resources, tools, and approaches. Plant Commun. **3**: 100323.

Macosko, E.Z. et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell **161**: 1202–1214.

Oughtred, R. et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. Publ. Protein Soc. **30**: 187–200.

Persson, S., Wei, H., Milne, J., Page, G.P., and Somerville, C.R. (2005). Identification of

genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc. Natl. Acad. Sci. U. S. A. **102**: 8633–8638.

Proost, S. and Mutwil, M. (2018). CoNekT: An open-source framework for comparative genomic and transcriptomic network analyses. Nucleic Acids Res. **46**: W133–W140.

Radivojac, P. et al. (2013). A large-scale evaluation of computational protein function prediction. Nat. Methods **10**: 221–227.

Rhee, S.Y. and Mutwil, M. (2014). Towards revealing the functions of all genes in plants. Trends Plant Sci. **19**: 212–221.

Richardson, L.G.L. and Schnell, D.J. (2020). Origins, function, and regulation of the TOC–TIC general protein import machinery of plastids. J. Exp. Bot. **71**: 1226–1238.

Ruprecht, C., Proost, S., Hernandez-Coronado, M., Ortiz-Ramirez, C., Lang, D., Rensing, S.A., Becker, J.D., Vandepoele, K., and Mutwil, M. (2017). Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. Plant J. **90**: 447–465.

Shibata, M. and Sugimoto, K. (2019). A gene regulatory network for root hair development. J. Plant Res. **132**: 301–309.

Stengel, A., Benz, J.P., Buchanan, B.B., Soll, J., and Bölter, B. (2009). Preprotein Import into Chloroplasts via the Toc and Tic Complexes Is Regulated by Redox Signals in Pisum sativum. Mol. Plant **2**: 1181–1197.

Szklarczyk, D. et al. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res.

Wang, J.-H., Zhao, L.-F., Wang, H.-F., Wen, Y.-T., Jiang, K.-K., Mao, X.-M., Zhou, Z.-Y., Yao, K.-T., Geng, Q.-S., Guo, D., and Huang, Z.-X. (2020). GenCLiP 3: mining human genes' functions and regulatory networks from PubMed based on co-occurrences and natural language processing. Bioinformatics **36**: 1973–1975.

Xiao, R., Zhang, C., Guo, X., Li, H., and Lu, H. (2021). MYB Transcription Factors and Its Regulation in Secondary Cell Wall Formation and Lignin Biosynthesis during Xylem Development. Int. J. Mol. Sci. **22**: 3560.

Yilmaz, A., Mejia-Guerra, M.K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS: The arabidopsis gene regulatory information server, an update. Nucleic Acids Res. **39**: D1118-22.

Yu, Y., Hu, H., Doust, A.N., and Kellogg, E.A. (2020). Divergent gene expression networks underlie morphological diversity of abscission zones in grasses. New Phytol. **225**: 1799–1815.