scientific reports

OPEN

An NLP-based method to mine gene and function relationships from published articles

Nilesh Kumar[1] & M. Shahid Mukhtar[1,2]✉

Understanding the intricacies of genes function within biological systems is paramount for scientific advancement and medical progress. Owing to the evolving landscape of this research and the complexity of biological processes, however, this task presents challenges. We introduce PATHAK, a natural language processing (NLP)-based method that mines relationships between genes and their functions from published scientific articles. PATHAK utilizes a pre-trained Transformer language model to generate sentence embeddings from a vast dataset of scientific documents. This enables the identification of meaningful associations between genes and their potential functional annotations. Our approach is adaptable and applicable across diverse scientific domains. Applying PATHAK to over 17,000 research articles focused on *Arabidopsis thaliana*, we assigned approximately 1493 GO terms to 10,976 genes by analyzing article sentences, comparing their embeddings to GO term embeddings, and mapping potential matches. The model demonstrates moderate-to-high predictive accuracy, capturing ~ 57% overlap of GO terms (6258 out of 10,976) between predicted and known annotations on TAIR, including 1271 and 161 exact matches and 4826 partially related terms. This method promises to significantly advance our understanding of gene functionality and potentially accelerate discoveries in the context of plant development, growth and stress responses in plants and other systems.

Keywords NLP, Gene function, Published articles, Gene ontology

Functional annotation is a challenge in the field of biology, as it involves identifying the functions of genes, proteins, and other biological entities[1,2]. Accurate functional annotation is important for understanding the mechanisms at work within biological systems and for predicting the effects of changes to these systems[3]. However, functional annotation can be difficult for several reasons. One reason is that the functions of many genes and proteins are not yet fully understood. In some cases, the functions of these entities may be poorly conserved across species, making it difficult to predict their functions in a given organism[4]. Additionally, the functions of numerous genes and proteins are context-dependent, highlighting that they may vary depending on the specific cellular or physiological context in which they are expressed[5]. This can make it difficult to accurately annotate the functions of these entities. Finally, the increasing amount of available biological data can make it challenging to accurately annotate the functions of all genes and proteins promptly.

Several methods can be used to annotate the functions of proteins or genes, including experimental approaches, computational approaches, hybrid approaches, and functional genomics approaches[6–8]. Experimental approaches involve conducting experiments to directly study the functions of genes and proteins, such as knock-out studies or overexpression studies[9,10]. Computational approaches involve using computer algorithms and databases to predict the functions of genes and proteins based on their sequences or other characteristics. Hybrid approaches involve combining experimental and computational approaches, while functional genomics approaches involve studying the functions of groups of genes rather than individual genes[11]. Accurate functional annotation is important for understanding the mechanisms at work within biological systems and for predicting the effects of changes to these systems. With technological advancement, the number of gene functions has increased dramatically in recent times.

Moreover, as technology has advanced, the ability to identify and study the functions of genes has also increased significantly. This has led to a rapid increase in the number of known gene functions in recent years. The development of techniques such as high-throughput sequencing and microarray analysis has allowed researchers to study the functions of large numbers of genes simultaneously, leading to a greater understanding of the roles that these genes play in biological processes[12]. Additionally, the availability of large amounts of

[1]Department of Biology, University of Alabama at Birmingham, 3100 East Science Hall, 902 14th Street South, Birmingham, AL 35294, USA. [2]Department of Genetics & Biochemistry, Clemson University, 206 Biosystems Research Complex, 105 Collings St., Clemson, SC 29634-0318, USA. ✉email: mshahid@clemson.edu

genomic data and the use of computational approaches to analyze this data has facilitated the identification of new gene functions. Overall, the rapid advancement of technology has greatly increased our understanding of gene functions and has opened up new avenues for research in this field. There have been numerous research articles published that focus on the functional annotation of genes and reading and understanding all of these articles can be a difficult and time-consuming task. This is due in part to the large number of articles that have been published on this topic, as well as the technical nature of the material covered in these articles[13]. To fully understand the functional annotation of genes, it is often necessary to have a strong background in molecular biology and genetics, as well as a familiarity with the methods and techniques used in functional annotation studies. Additionally, the complex nature of biological systems and the interactions between genes and their functions can make it challenging to fully understand the implications of functional annotation findings.

Attempts have been made to understand the contents of research articles using document-level representation techniques. The "Document-level representations of scientific articles" refer to a vector representation of a scientific article that captures the meaning or content of the entire article as a whole[14]. These representations are used in various natural language processing (NLP) tasks, such as information retrieval, recommendation, and classification. The goal of generating document-level representations is to create a compact and useful representation of the article that can be used in downstream NLP tasks. For example, in information retrieval, the representation can be used to measure the similarity between a query and a set of documents and return the most relevant documents. In classification, the representation can be used as an input to a machine learning model to train it to classify the article into different categories[15,16].

Document-level representations can be generated using a variety of techniques, including latent semantic analysis (LSA)[17], latent dirichlet allocation (LDA)[17,18], Doc2Vec[18], and transformer-based models like BERT[15,16,19], GPT-2[20], and SPECTER[14]. The latent semantic analysis (LSA) model uses linear algebra to identify latent concepts within a set of documents and create a low-dimensional representation of each document in terms of those concepts. The latent dirichlet allocation (LDA) uses probabilistic generative methods to identify latent topics within a set of documents and create a low-dimensional representation of each document in terms of those topics. Doc2Vec uses neural networks to create a high-dimensional vector representation of each document based on the words it contains. Transformer-based models, such as BERT and GPT-2, are pre-trained using a large amount of text data and can be fine-tuned to create representations of specific types of documents, such as scientific articles.

SPECTER also uses transformer architecture and a citation network encoder that creates a document-level representation of scientific articles with the help of citation structure between articles. Overall, these methods create a vector representation of a scientific article by capturing different aspects of the article, such as the words it contains, the topics it covers, or the relationships between articles.

Here, we developed PATHAK, a purpose-built language model for mining gene-function relationships from scientific literature. It has several advantages over ChatGPT and Google GEMINI. First, PATHAK is specifically optimized for this task, while ChatGPT is a general-purpose language model and Google GEMINI primarily focuses on data access and analysis. This allows PATHAK to perform a more systematic and accurate analysis of research articles, resulting in more reliable gene-function relationship predictions. Second, PATHAK employs a structured approach to mitigate hallucination issues that can be prevalent in large language models. By leveraging sentence embeddings and semantic similarity scores, PATHAK ensures that the assigned GO terms are based on meaningful associations between genes and the context of the sentences. This approach enhances precision and reduces the likelihood of hallucinations. ChatGPT and Google GEMINI, on the other hand, may not employ specific methods to address hallucination issues, potentially leading to inaccurate results. Overall, PATHAK's specialized focus and structured methodology offer a more reliable approach to gene-function relationship mining compared to ChatGPT and Google GEMINI.

Related work

The Relation Extraction allows gene–disease associations to be extracted from the literature. OnTheFly2.0 is one such tool, and it is a text-mining application that recognizes biomedical entities automatically. It does document annotation, network analysis, and functional enrichment analysis. It allows users to extract relevant information from large amounts of text data and then analyze it to identify patterns and relationships[21].

RENET2 is another highly accurate and efficient tool and can handle large amounts of data. One of the critical features of RENET2 is that it uses an iterative training data expansion to improve the model's performance over time. This allows the system to continually learn and improve its ability to extract gene-disease relations from text[22].

Like RENET2, Darling is based on a machine learning technique that uses text data from scientific publications to identify patterns and relationships between different biomedical entities. The application allows users to search for specific diseases, genes, proteins, etc., and returns a list of associations that have been identified in the literature[23].

To be able to mine hidden gene functions from a huge number of research articles, we propose PATHAK, a method for predicting the functions of genes based on sentence-level embeddings of scientific documents. It uses a pre-trained Transformer language model, a type of neural network architecture commonly used for natural language processing tasks, to generate embeddings of sentences within research articles and embeddings of standard Gene Ontology (GO) term annotations.

These embeddings are then used to link genes with their potential functions or GO terms. The approach is based on the idea that the functions of genes are often described in the scientific literature, and by analyzing the language used in these documents, it is possible to infer the functions of genes based on the contexts in which they are mentioned. Overall, PATHAK is a promising method for predicting gene functions and may be useful for a variety of applications in molecular biology and genetics.

Using PATHAK, we analyzed over 17,000 PubMed research articles related to *A. thaliana* and its associated genes. Our analysis using PATHAK resulted in the confirmation of GO ontologies (TAIR[24]) for approximately 700 genes.

Overall, the output of the PATHAK pipeline is a list of gene-function relationships, each of which is supported by a set of relevant sentences from the scientific literature. This information can be used by researchers to deepen their understanding of gene functionality and to identify new and promising avenues for research.

Methods

The tutorial of PATHAK workflow along with code is available on GitHub (https://nilesh-iiita.github.io/PATHAK/intro.html). PATHAK workflow is a systematic process that can be broken down into four key steps. To begin, we analyze research articles and identify paragraphs that contain known genes. These paragraphs are then further segmented into individual sentences. Additionally, we compile a list of definitions for all possible Gene Ontology (GO) terms. The next step involves dividing the multi-line groups of sentences that make up each GO term definition into individual sentences and assigning a relevant GO term to each one. Afterward, we use embedding techniques to represent the sentences from research articles and GO term definitions as numerical vectors. Lastly, we compare the embeddings of the GO terms and article sentences to accurately assign potential GO terms to each sentence (Fig. 1).

Research Article curation

In this study, we obtained a total of 5,108,849 research articles in extensible markup language (XML) format from PubMed. Then, by utilizing keywords such as *A. thaliana* gene names, "Plant", and "arabidopsis, thaliana[25], *A. thaliana*", we narrowed down the research articles to a total of 82,000 potential candidates. However, for this study, we only analyzed a total of 17,322 articles (Supplementary Table 1, "List Of PMC IDs").

With the help of the Beautiful Soup Python package, the XML articles were parsed and broken down into paragraphs and sentences.

GO definitions

Gene ontology (GO) is a controlled vocabulary that provides a standardized way of describing the properties and functions of genes and gene products across all species. It is widely utilized in bioinformatics and systems biology to annotate and analyze genomic data. The GO is divided into three main branches: Cellular components (CC), which describe the location of gene products within a cell, Molecular Function ontologies (MF), which describe the molecular activities of gene products, and Biological Process ontologies (BP), which describe the overarching biological processes that a gene product participates in. In this work, we focused solely on the BP ontology.

The GO is organized in a hierarchical structure, where each term is arranged in a directed acyclic graph (DAG) format. This structure comprises nodes, which represent the terms in the ontology, and edges, which depict the relationships between the terms. Each term in the ontology is assigned a unique identifier, such as an "accession number" or "GO ID", along with additional information such as the ontology it belongs to (e.g. biological_process), definition, a name that describes the term's meaning, and synonyms, alternative names for the term.

Sentence embedding is a technique for encoding textual data as fixed-length vectors. It is based on the idea of word embedding, which maps each word in a given corpus to a numerical vector that captures the semantic relation between words. However, sentence embedding can represent longer pieces of text, such as entire sentences, in numerical form.

By using SentenceTransformers to generate embeddings for the sentences in the research articles and the GO term definitions, we were able to obtain a numerical representation of the sentences that captures their meaning and context, which can be easily used to sentence similarity tasks. SentenceTransformers is a library that provides pre-trained models for sentence embedding generation. These models are trained on large corpora of text data and can be fine-tuned on specific tasks or domains. The "sentence_embedding_biobert" ("https://sparknlp.org/2020/09/19/biobert_pubmed_base_cased.html") pre-trained model was used to get the embeddings, which is a pre-trained language BERT (Bidirectional Encoder Representations from Transformers) model developed specifically for biomedical text.

We used this similarity score to measure how similar or dissimilar two sentences are. By comparing the embeddings of the article sentences GO term definitions and calculating a similarity score, the authors were able to identify the sentences that are semantically similar and can be used to assign a potential GO term to a sentence of a paragraph that contains at least a gene, in a research article. This is a crucial step as it enabled us to accurately assign GO terms to sentences in a research article, which can help in understanding the biological processes that are discussed in the article.

To get the similarity score the "cosine_similarity" function was used from the Sklearn python library (v.1.1.3). We empirically determined the 0.92 cosine-similarity threshold by examining a range of candidate cutoffs (e.g., 0.92–0.98) on a validation set. We compared how each threshold affected our main metrics (Precision, Recall, F1, and Jaccard) and found that 0.92 yielded the best balance between including truly relevant sentence pairs and minimizing false-positive assignments. Additionally, the performance table (see Supplementary Table 1) shows that similarity scores typically cluster around 0.92–0.95 for correct matches, making 0.92 a natural choice to maximize predictive accuracy. This threshold thus reflects both empirical testing of various cutoffs and the observed distribution of similarity scores.

$$\cos(\theta) = \frac{a \cdot b}{|a|\,|b|} \tag{1}$$

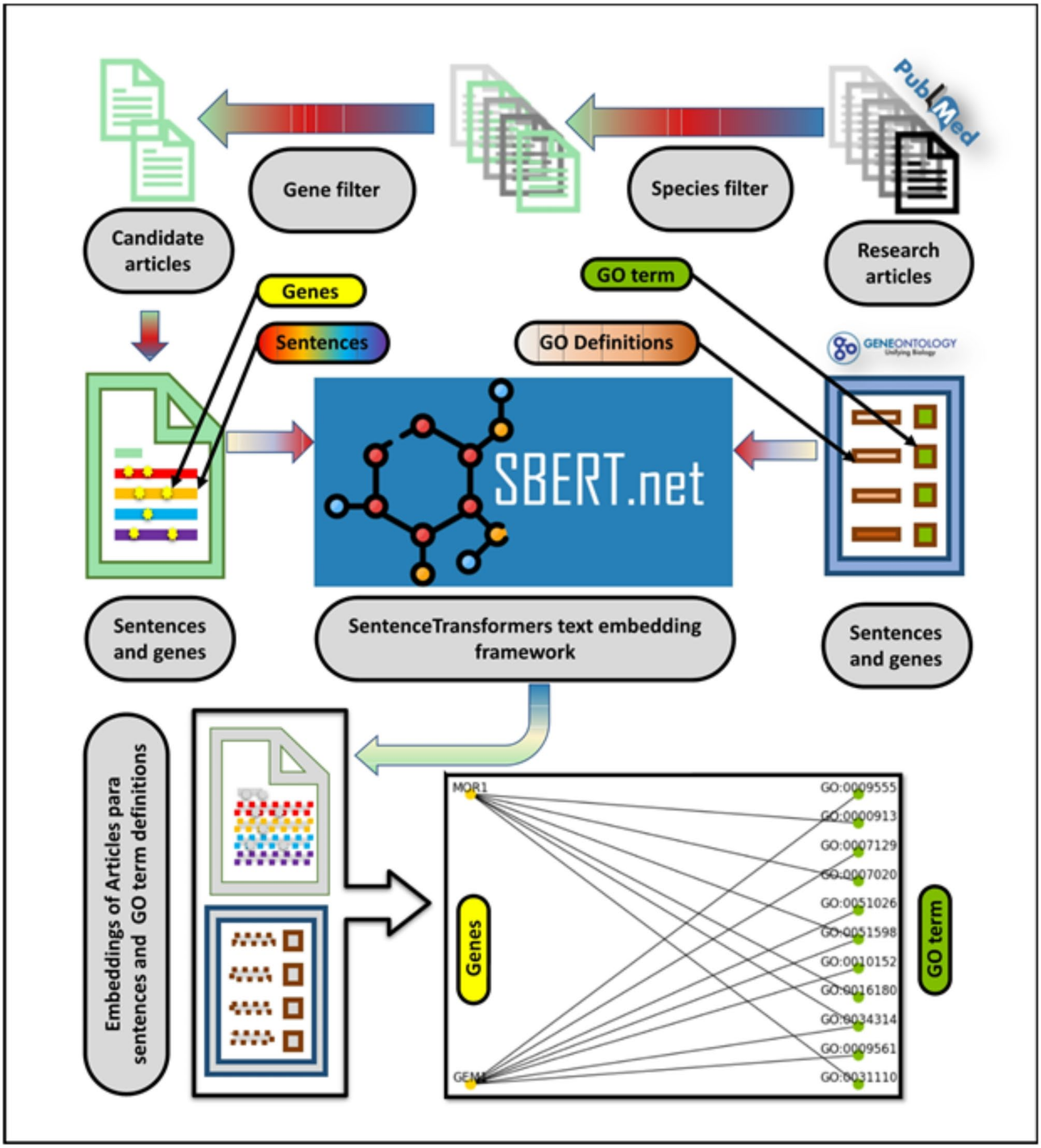

Fig. 1. PATHAK pipeline for mining gene-function relationships from scientific literature. The pipeline consists of the 5 steps. (i) Candidate article selection: PATHAK uses a gene filter and a species filter to select a subset of candidate articles from a repository of research articles. (ii) Sentence extraction: PATHAK extracts sentences from the candidate articles, focusing on sentences that contain genes. (iii) Sentence embedding: PATHAK generates sentence embeddings for the extracted sentences using a pre-trained sentence transformer model. (iv) GO term prediction: PATHAK predicts the GO terms associated with each sentence using a machine learning model that is trained on a dataset of sentence embeddings and GO terms. (v) Gene-function relationship mining: PATHAK integrates the predicted GO terms for each sentence to identify gene-function relationships.

Where,

a and b are vectors, e.g. GO and sentence of the article paragraph.

θ is the angle between the two vectors.

||a|| and ||b|| are the magnitudes of the vectors a and b, respectively.

GO term assignment

When a paragraph includes only one gene, "G1" and a single sentence "S1" that is found to be similar to a single GO term definition "GO1" through cosine similarity, then the gene "G1" will be associated with the GO term "GO1". This simple approach ensures that the gene is assigned the most relevant GO term based on the sentence's semantic similarity.

However, in a situation where there are multiple genes in the paragraph (G1, G2,.Gn) and a single sentence "S1" that is found to be similar to a single GO term definition "GO1" through cosine similarity, then all genes (G1, G2,.Gn) will be associated with the GO term "GO1". This approach ensures that all the genes in the paragraph are assigned the most relevant GO term based on the sentence's semantic similarity.

Similarly, if there are multiple (n) gene in the paragraph (G1, G2,.Gn) and multiple (s) sentence "S1,S2,…Ss" that is found to be similar to multiple (g) single GO term definitions "GO1, GO2,.Gog" through cosine similarity, then the all 'n' genes will be associated with the "g" GO terms. This method ensures that all the genes in the paragraph are assigned the most relevant GO terms based on the semantic similarity of the sentences.

Data availability

In this research endeavor, our initial data collection phase involved sourcing an extensive repository of scholarly literature. We accessed a vast corpus of research articles amounting to a substantial total of 5,108,849 documents, all formatted in extensible markup language (XML), obtained from the reputable PubMed database. To refine this extensive dataset and hone in on the pertinent research related to *A. thaliana* gene names and plant biology, we employed targeted keywords including *A. thaliana* gene names, "Plant," and "arabidopsis, thaliana", "*A. thaliana*". These strategic keywords helped us filter and narrow down the pool of research articles to a more manageable subset of 82,000 potential candidates, aligning closely with the specific focus of our study. However, to maintain a rigorous and comprehensive analysis, we selected a meticulously curated subset of 17,322 articles for in-depth examination and analysis, ensuring a thorough investigation into the intricate domain of gene-function relationships in the context of plant biology.

Results

We employed PATHAK to analyze 17,322 research articles that provided valuable insights into the biological processes discussed by associating a large number of genes with relevant Gene Ontology (GO) terms (Supplementary Table 1, "List Of PMC IDs"). We used a method involving breaking down articles into paragraphs and sentences, listing definitions of all possible GO terms, comparing sentence embeddings with GO term embeddings to assign a potential GO term to each sentence, and finally associating 10,976 genes with about 1,493 GO terms (Fig. 2). This set of genes are designated as "predicted". The results of this method, known as PATHAK, were validated using existing GO annotations from the TAIR database[24]. It is important to note that this work is focused specifically on Biological Process (BP) ontologies. TAIR employs a combination of manual and computational techniques to generate Gene Ontology (GO) annotations, resulting in a total of 24,283 available locus annotations. Out of the 10,976 predicted genes, 10,473 represent the subset for which TAIR possesses at least one gene function association, and we defined them as "known". Comparative analyses revealed that 1,432 genes within both the predicted and known categories share identical annotations and are labelled as "common" (Fig. 2 and Supplementary Table 1). Considering the involvement of genes in various biological processes, it is expected that a single gene can be associated with multiple GO terms. Similarly, PATHAK may predict more than one GO term for a single gene. To address this, we extended our analysis to GO tree data structure encompassing descendants and ancestors. The resulting combined list of dependents and ancestors is denoted as 'relative GO terms', whereas the comparative analysis between PATHAK "predicted" and TAIR "known" gene lists revealed 6097 common genes that are referred to as "Common relative GO terms" (Fig. 2 and Supplementary Table 1). These 'Common relative GO terms' hold significant importance in the annotation process, as they enable the inference of novel gene functions for such loci that already possess known GO annotations. Overall, comparative study between PATHAK "predicted" and TAIR "known" genes identified 6258 genes (57%) that are common and possess at least one "common relative" (Fig. 2a), highlighting the biological validation of our study. As for the remaining 4,718 genes between PATHAK and TAIR gene lists, it's worth noting that PATHAK also generated additional GO term predictions that are not found in TAIR. Finally, PATHAK exclusively predicted annotations for 503 genes, which accounts for approximately 5%, as these annotations are not present in TAIR. Taken together, this data underscores the significance of PATHAK in enhancing the prediction of GO terms.

Additionally, to further analyze the sentence embeddings, dimension reduction techniques were applied using the UMAP algorithm with the aid of the umap-learn python library (v.0.5.3). Both supervised and unsupervised UMAP dimension reduction analyses were conducted on the top 10 GO terms that were assigned to the majority of the genes (Fig. 2b, c). This approach allowed for an in-depth examination of the relationships and similarities between the sentence embeddings and the assigned GO terms, and it provided a more comprehensive understanding of the biological processes discussed in the research articles.

Evaluation of the accuracy or performance of a model designed for complex data, such as GO term annotations, is inherently challenging. Assuming TAIR as the ground truth, it contains annotations at varying

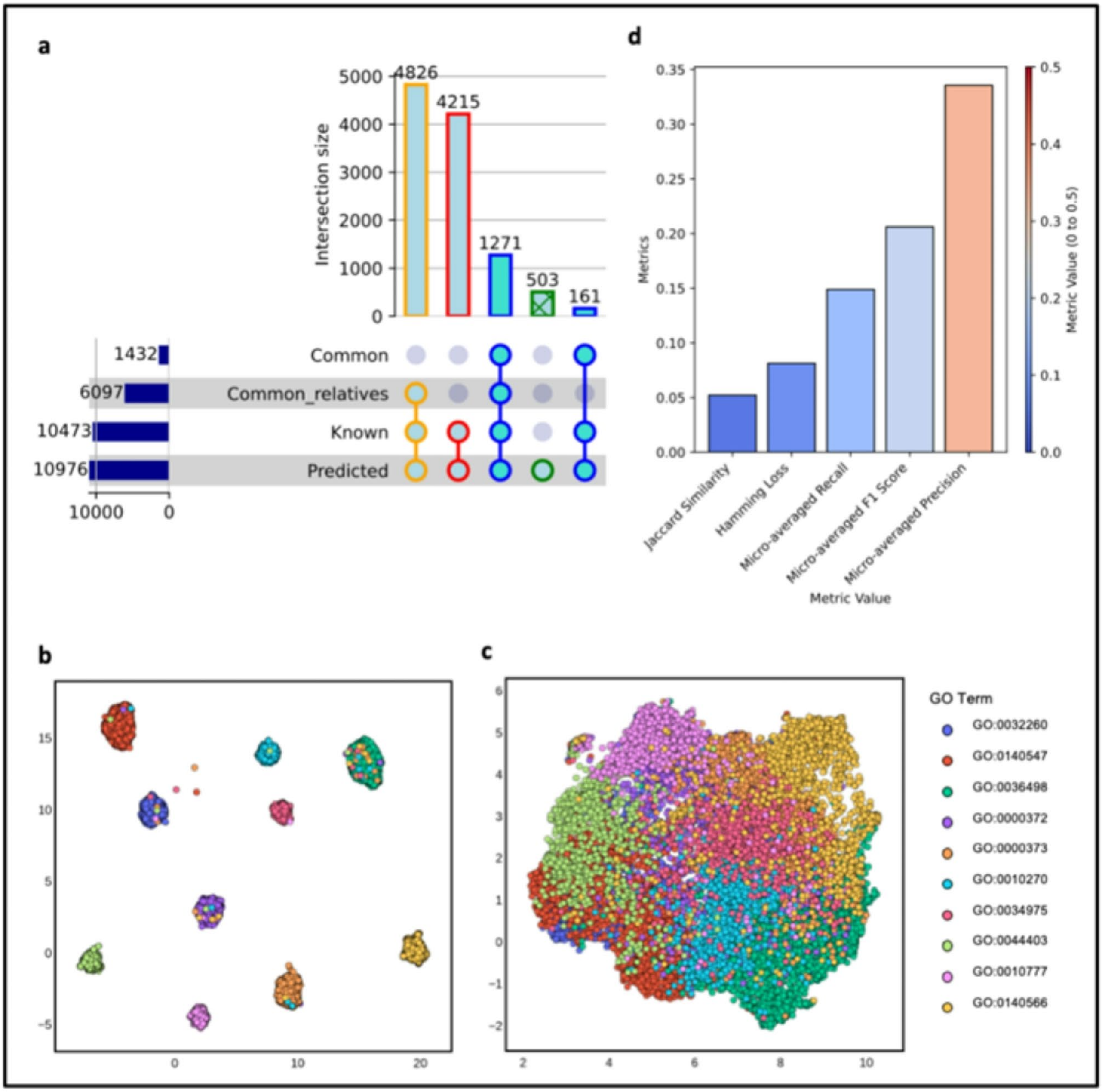

Fig. 2. UpSet plot illustrating the overlap between predicted and known GO terms for *Arabidopsis thaliana*. The horizontal bars represent the total size of each category: 'Predicted' (GO terms predicted by the model), 'Known' (ground-truth GO annotations), 'Common_relatives' (related GO terms predicted), and 'Common' (exact GO terms correctly predicted). The vertical bars display the intersection sizes between these categories, highlighting the number of GO terms shared among them (**a**). Supervised and unsupervised UMAP dimension reduction analyses on embedding of sentences found like definition of the top 10 GO terms that were assigned to the majority of the genes (**b**,**c**). Performance evaluation metrics for GO term prediction. The bar plot represents various metrics, including Jaccard Similarity, Hamming Loss, Micro-averaged Recall, Micro-averaged F1 Score, and Micro-averaged Precision, based on the comparison between predicted and known GO terms. Metric values are color-coded, with a gradient from blue to red indicating lower to higher values, respectively (**d**).

levels of specificity, which can complicate direct comparisons. To fairly evaluate the model's performance, the predicted GO terms from PATHAK and the TAIR annotations were standardized to represent the same level of annotation by aligning them using the hierarchical structure of the GO annotation tree from top to bottom. The performance metrics, as shown in Fig. 2d, highlight the model's effectiveness. The model achieved a micro-averaged precision of ~ 0.35, a recall of ~ 0.2, and an F1 score of ~ 0.25, indicating moderate accuracy in predicting GO terms. The Jaccard similarity index was ~ 0.1, reflecting the overlap between predicted and known

GO terms, while the Hamming loss was ~ 0.15, indicating a low level of label misclassification. These results demonstrate that PATHAK can effectively predict GO annotations, even when faced with the complexities of multi-label classification and hierarchical relationships.

Case study

To evaluate and understand the predictions by PATHAK one research article by de Jager, Nicholas, et al.[26] is chosen that focuses on an important micronutrient sulfur[27]. Most of the prediction were done PATHAK was on fourth paragraph of the "Discussion" section. In that paragraph, the authors point out that CAD2 is a particular mutant allele that disrupts the *GSH1* gene in *Arabidopsis thaliana*. The *GSH1* gene encodes the rate-limiting enzyme in glutathione (GSH) synthesis, so a *CAD2* (*gsh1*) mutant is compromised in its ability to produce GSH. The paragraph specifically notes that *cad2* plants are "hyposensitive" to phosphate deprivation, implying that altered GSH levels (due to the defect in GSH1) affect the plant's response to low phosphate conditions. In other words, *GSH1* is critical for glutathione biosynthesis, while *CAD2* is the name of a mutant allele of that gene, and together they influence how Arabidopsis manages phosphate stress via changes in GSH content.

Summary of results by PATHAK for *CAD2* or *GSH1* (AT4G23100)

1. Sentence:
2. "GSH is part of the redox system preventing cell damage, by reversibly undergoing redox reactions at its thiol group, and thus maintaining the redox status of the cell (Noctor et al., 2012)."

- Predicted GO Term: GO:0006749 (glutathione metabolic process).
- Score: 0.9426.
- Depth: 3.
- Correctness: This prediction is correct. The sentence explicitly describes the role of glutathione (GSH) in redox reactions, which aligns directly with the "glutathione metabolic process."

3. Sentence:
4. "However, GSH and GSLs are tightly connected, as GSH provides reduced S for GSL synthesis, and can indirectly affect GSL synthesis (Geu-Flores et al., 2009)."

- Predicted GO term: GO:0010142 (farnesyl diphosphate biosynthetic process, mevalonate pathway).
- Score: 0.9321.
- Depth: 6.
- Correctness: This prediction is incorrect. The sentence discusses the role of GSH in providing reduced sulfur for glucosinolate (GSL) synthesis. Farnesyl diphosphate biosynthesis (GO:0010142) is unrelated to this context, as it pertains to sterol biosynthesis rather than sulfur metabolism.

5. Sentence:
6. "GSH levels are known to follow diurnal and seasonal cycles in some plant species, and high levels of GSH in the winter are linked to winter hardiness and protection against winter injury (Turunen and Latola, 2005)."

- Predicted GO term: GO:0140547 (acquisition of seed longevity).
- Score: 0.9341.
- Depth: 2.
- Correctness: This prediction is partially incorrect. While the sentence emphasizes GSH's role in protecting plants during winter, the "acquisition of seed longevity" GO term is only loosely related. The context is more about GSH's role in abiotic stress tolerance rather than seed longevity.

7. Sentence:
8. "Although phosphate deprivation does not affect total GSH content in plants, it increases the reduced to oxidized ratio (GSH/GSSG) (Trujillo-Hernandez et al., 2020)."

- Predicted GO term: GO:0051775 (response to redox state).
- Score: 0.9276.
- Depth: 2.
- Correctness: This prediction is correct. The sentence explicitly discusses the redox state balance (GSH/GSSG ratio), which directly relates to the "response to redox state" GO term.

9. Sentence:
10. "Our results suggest that GSH and GSLs are regulated in an opposite way; namely, accessions with higher S and sulfate content, higher GSL content, and higher sulfate uptake had lower GSH content and also lower phosphate anions."

- Predicted GO term: GO:0010264 (myo-inositol hexakisphosphate biosynthetic process).
- Score: 0.9247.
- Depth: 5.
- Correctness: This prediction is incorrect. The sentence discusses the relationship between GSH, GSLs, and sulfate metabolism, which is unrelated to the "myo-inositol hexakisphosphate biosynthetic process." This prediction likely stems from the mention of "phosphate anions," but the context is incorrect.

Overall evaluation of PATHAK results for this case study

- Correct predictions:
  - GO:0006749 (glutathione metabolic process).
  - GO:0051775 (response to redox state).
- Partially correct prediction:
  - GO:0140547 (acquisition of seed longevity).
- Incorrect predictions:
  - GO:0010142 (farnesyl diphosphate biosynthetic process, mevalonate pathway).
  - GO:0010264 (myo-inositol hexakisphosphate biosynthetic process).

Is the result correct?

Accuracy assessment

- Out of 5 predictions, 2 are correct, 1 is partially correct, and 2 are incorrect.
- While PATHAK demonstrates an ability to predict GO terms relevant to the provided context (e.g., GO:0006749 and GO:0051775), it struggles with more complex or indirect relationships (e.g., the connection between GSH and GSLs).

Reasoning behind correctness and errors

1. Correct predictions: These are based on explicit matches between the sentence context and the GO term definition. For example, the mention of "redox system" aligns directly with the "response to redox state" GO term.
2. Partially correct prediction: When the context is loosely related to the GO term, such as GSH's seasonal cycles being interpreted as "seed longevity," PATHAK appears to overgeneralize based on limited keywords.
3. Incorrect predictions: Errors arise when PATHAK misinterprets specific terms (e.g., "phosphate anions" leading to "myo-inositol hexakisphosphate biosynthetic process") or fails to capture the nuanced context of sulfur and glucosinolate metabolism.

Overall PATHAK conclusions for this case study
PATHAK demonstrates moderate predictive accuracy for gene function prediction, particularly for direct and explicit relationships between sentences and GO terms. However, incorrect predictions highlight limitations in handling complex or indirect associations, which may require incorporating domain-specific ontologies or additional contextual data for refinement.

Discussion

The task of understanding gene function can be complex and challenging, but it is crucial for advancing our understanding of biological systems and developing new treatments for diseases. There are several reasons for this complexity, including the multiple mechanisms and pathways involved, the constantly evolving nature of our understanding, the need for a foundation in biology and related fields, and the specialized language used in research articles. To address these challenges, we have proposed a new method called PATHAK, which uses a pre-trained Transformer language model to link genes with their potential function by analyzing sentence-level embeddings of scientific documents. This method is flexible and applicable to a wide range of scientific domains and has the potential to advance our understanding of gene function and facilitate the discovery of new treatments for diseases. It was applied to a large dataset of articles focusing on *A. thaliana* and the genes related to this plant species, which adds to our understanding of the complex mechanisms at work within this and similar plant species.

There are several research articles already published which include several aspects of gene functions, but the information is not presented in a standardized format, making it ambiguous and difficult to understand. The lack of standardization makes it challenging for researchers to easily access and compare the information across different studies. This can lead to confusion and inconsistencies in the understanding of gene functions, slowing down progress in the field.

Furthermore, even when gene functions are explicitly discussed and mentioned in a research article, they are often not updated in standard oncology databases such as Gene Ontology (GO). This can lead to inaccuracies and inconsistencies in the information available to researchers, making it difficult to have a comprehensive understanding of gene functions. This also makes it challenging for researchers to keep track of the latest findings and developments in the field.

Reading research articles can also be a cumbersome task for researchers. The articles are often written in a specialized language and may contain complex technical details, which can make it difficult for non-experts to understand and extract the relevant information on gene functions. This can be a time-consuming process and can limit the ability of researchers to keep up with the latest findings and developments in the field.

All these challenges make it difficult to have a comprehensive understanding of gene functions, and this is where the PATHAK method comes into play. It tries to overcome these challenges by providing a way to link genes with the potential function using sentence-level embedding of scientific documents based on pre-training a Transformer language model, which is accessible and easy to use, this can greatly help researchers in understanding the gene functions and also help in updating the standard databases like GO.

For our study, we selected *A. thaliana*, also known as *thale cress*, as our model organism. *A. thaliana* has several genetic and molecular tools available for studying its biology, such as T-DNA insertion lines, which allow for the study of specific genes, and a variety of mutant lines that have been generated to study different aspects of plant growth and development. Moreover, *A. thaliana* shares many similarities with other plants, including crop plants, which makes it a useful model for understanding the biology of other plants. This makes it a great model organism for studying plant growth, development, and response to environmental stress.

Our proposed method, PATHAK, was used to analyze over 17,000 research articles that contained keywords related to *A. thaliana*, as well as related gene symbols or IDs. By doing so, we were able to assign gene functions to over 10,000 genes. Among these, 1,493 GO annotations by PATHAK correlated with TAIR GO annotations and around 6000 gene in "common relative" category, exhibiting overall ~ 57% accuracy. However, for the remaining annotations most of them have confounding gene symbol or IDs, for which further manual verification is necessary (Supplementary Table 1, "Genes with confounding Aliases"). While this process does require some additional effort, it is a relatively quick and straightforward task in comparison to reading through all of the articles in their entirety. The model evaluation reveals that PATHAK achieved a micro-averaged precision of ~ 0.35 and a recall of ~ 0.2, resulting in an F1 score of ~ 0.25. These values indicate moderate predictive performance. The Jaccard similarity index was ~ 0.1, reflecting the overlap between the predicted and known GO terms, while the Hamming loss was ~ 0.15, indicating a relatively low rate of label misclassification.

These findings suggest that PATHAK demonstrates reasonable accuracy in predicting GO annotations, particularly given the complexity of the hierarchical relationships and the multi-label nature of the data. The moderate precision and recall highlight the method's ability to balance specificity and sensitivity, which is critical for applications requiring accurate functional annotations of genes. Furthermore, these results underscore the importance of aligning annotation levels to facilitate robust model evaluation in future studies.

As detailed in Supplementary Text 1, To illustrate PATHAK's performance in a real-world context, we analyzed its predictions on a single research article exploring *Arabidopsis thaliana* genes CAD2 and GSH1. Of five GO terms predicted, two (glutathione metabolic process; response to redox state) were correct, one (acquisition of seed longevity) was partially correct, and two (farnesyl diphosphate biosynthetic process; myo-inositol hexakisphosphate biosynthetic process) were incorrect. Inaccuracies generally stemmed from keyword overlaps (e.g., "phosphate anions" misinterpreted as "myo-inositol hexakisphosphate biosynthetic process") or incomplete context capture (e.g., conflating GSH's role in winter hardiness with seed longevity).

Another limitation of the current study lies in the absence of a comparative analysis between functional annotations derived from protein sequences and those obtained through literature mining using PATHAK. Functional annotations based on protein sequences, such as those generated by tools like BLAST2GO and InterProScan, provide insights grounded in evolutionary and structural information, while literature-mined annotations from PATHAK offer a textual, context-driven perspective rooted in existing research. These two approaches have the potential to yield complementary insights. In the future studies, however, this comparative framework would offer a more comprehensive understanding of gene function by leveraging the strengths of both approaches, facilitating the discovery of novel biological insights and improving annotation reliability in the context of complex genomic datasets.

Despite these limitations, PATHAK effectively identified explicit associations, especially when the sentence content directly matched known GO term definitions (e.g., redox-related processes). These findings reveal that PATHAK excels with clear, direct statements but can struggle when complex biological relationships or subtle context are involved. PATHAK can be applicable to newly sequenced genomes and models such as duckweed[28]. Future refinements such as integrating domain-specific ontologies or leveraging additional textual cues—may help reduce these incorrect or partially correct assignments, further enhancing PATHAK's utility in gene function prediction.

One of the disadvantages of the PATHAK method is that it does not guarantee that article that is being processed is specifically written for a particular task. For example, an article may be written for any crop plant but due to the versatility of *A. thaliana* of keywords related to *A. thaliana* are mentioned in almost every research article. Even some genes also overlap with animals. This problem can be overcome by selecting research articles with some prior knowledge. Further, the Gene ID system is very crucial to understanding gene function. Whereas multiple gene systems are assigned to one TAIR ID (AGI) for *A. thaliana* and often research articles use gene symbols, not TAIR IDs.

PATHAK exhibits several advantages over both ChatGPT and Google GEMINI in the domain of mining gene-function relationships from scientific literature. Firstly, PATHAK is purpose-built to tackle a specific task: extracting meaningful relationships between genes and their functions from research articles. In contrast, ChatGPT is a general-purpose language model, not optimized for this specialized task, and Google GEMINI primarily focuses on data access and analysis rather than gene-function relationship mining. PATHAK's targeted approach allows for a systematic analysis of research articles and accurate assignment of potential Gene Ontology (GO) terms to sentences, thus providing researchers with a powerful tool to deepen their understanding of gene functionality.

In addressing the challenge of language model hallucination, PATHAK employs a structured and methodical approach that helps mitigate hallucination issues that can be prevalent in large language models (LLMs) like ChatGPT. By leveraging sentence embeddings and semantic similarity scores, PATHAK ensures that the assigned GO terms are based on meaningful associations between genes and the context of the sentences. This approach enhances precision and reduces the likelihood of hallucinations or inaccurate associations. On the other hand, ChatGPT, being a more flexible and open-ended language model, may generate responses that are not entirely accurate or contextually appropriate, potentially leading to hallucination. Google GEMINI, while useful for data access and analysis, may not employ specific methods to address hallucination issues related to language understanding and generation. Therefore, PATHAK's structured methodology offers a more controlled and reliable approach in the context of gene-function relationship mining compared to ChatGPT and Google GEMINI.

Data availability

The tutorial of PATHAK workflow along with code is available on GitHub (https://nilesh-iiita.github.io/PATHAK/intro.html).

Received: 14 April 2024; Accepted: 24 February 2025
Published online: 03 March 2025

References

1. Fakhar, A. Z., Liu, J. B., Pajerowska-Mukhtar, K. M. & Mukhtar, M. S. The ORFans' tale: new insights in plant biology. *Trends Plant Sci.* **28**, 1379–1390. https://doi.org/10.1016/j.tplants.2023.06.011 (2023).
2. Fakhar, A. Z., Liu, J. B., Pajerowska-Mukhtar, K. M. & Mukhtar, M. S. The lost and found: unraveling the functions of orphan genes. *J. Dev. Biol.* **11** https://doi.org/10.3390/jdb11020027 (2023).
3. Ejigu, G. F. & Jung, J. Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biol. Basel* **9**, 295. https://doi.org/10.3390/biology9090295 (2020).
4. Ponting, C. P. Biological function in the Twilight zone of sequence conservation. *BMC Biol.* **15**, 71. https://doi.org/10.1186/s12915-017-0411-5 (2017).
5. Buchberger, E., Reis, M., Lu, T. H. & Posnien, N. Cloudy with a chance of insights: context dependent gene regulation and implications for evolutionary studies. *Genes (Basel)* 10. https://doi.org/10.3390/genes10070492 (2019).
6. Pellegrini, M. Computational methods for protein function analysis. *Curr. Opin. Chem. Biol.* **5**, 46–50. https://doi.org/10.1016/s1367-5931(00)00165-4 (2001).
7. Benso, A. et al. A combined approach for genome wide protein function annotation/prediction. *Proteome Sci.* **11**, S1. https://doi.org/10.1186/1477-5956-11-S1-S1 (2013).
8. de Crecy-Lagard, V. et al. A roadmap for the functional annotation of protein families: a community perspective. *Database (Oxford)* https://doi.org/10.1093/database/baac062 (2022).
9. Ahmed, H. et al. Network biology discovers pathogen contact points in host protein-protein interactomes. *Nat. Commun.* **9**, 2312. https://doi.org/10.1038/s41467-018-04632-8 (2018).
10. Zhang, J., Zheng, N. & Zhou, P. Exploring the functional complexity of cellular proteins by protein knockout. *Proc. Natl. Acad. Sci. USA* **100**, 14127–14132. https://doi.org/10.1073/pnas.2233012100 (2003).
11. Singh, P., Mondal, S. & Singh, R. L. *In Advances in Animal Genomics* (eds Sukanta, M. & Ram L. S.,) 1–12 (Academic, 2021).
12. Hong, J. et al. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Brief. Bioinform.* **21**, 1437–1447. https://doi.org/10.1093/bib/bbz081 (2020).
13. Milošević, N. & Thielemann, W. Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *J. Web Semant.* **75**, 100756 (2023).
14. Cohan, A. et al. Document-level representation learning using citation-informed transformers. *arXiv* **2004**, 07180 (2020).
15. Lee, J. et al. BioBERT: a pre-trained biomedical Language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240. https://doi.org/10.1093/bioinformatics/btz682 (2020).
16. Bhasuran, B. BioBERT and similar approaches for relation extraction. *Methods Mol. Biol.* **2496**, 221–235. https://doi.org/10.1007/978-1-0716-2305-3_12 (2022).
17. Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. *Data Mining* (eds Ian H. Witten, Eibe Frank, Mark A. Hall, & Christopher J. Pal) 503–532 (Morgan Kaufmann, 2017).
18. Kim, D., Seo, D., Cho, S. & Kang, P. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Inf. Sci.* **477**, 15–29. https://doi.org/10.1016/j.ins.2018.10.006 (2019).
19. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. & Bert pre-training of deep bidirectional transformers for language understanding. *ArXiv* 181004805 (2018).
20. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog.* **1**, 9 (2019).
21. Baltoumas, F. A. et al. OnTheFly(2.0): a text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis. *NAR Genom. Bioinform.* **3**, lqab090. https://doi.org/10.1093/nargab/lqab090 (2021).
22. Su, J., Wu, Y., Ting, H. F., Lam, T. W. & Luo, R. RENET2: high-performance full-text gene-disease relation extraction with iterative training data expansion. *NAR Genom. Bioinform.* **3**, lqab062. https://doi.org/10.1093/nargab/lqab062 (2021).
23. Karatzas, E. et al. Darling: A web application for detecting disease-related biomedical entity associations with literature mining. *Biomolecules* **12**, 520. https://doi.org/10.3390/biom12040520 (2022).
24. Swarbreck, D. et al. The arabidopsis information resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–1014. https://doi.org/10.1093/nar/gkm965 (2008).
25. Parry, G., Provart, N. J., Brady, S. M. & Uzilday, B. Multinational Arabidopsis steering, C. Current status of the multinational Arabidopsis community. *Plant. Direct* **4**, e00248. https://doi.org/10.1002/pld3.248 (2020).
26. de Jager, N. et al. Traits linked to natural variation of sulfur content in Arabidopsis thaliana. *J. Exp. Bot.* **75**, 1036–1050. https://doi.org/10.1093/jxb/erad401 (2024).
27. Wang, W., Liu, J., Mishra, B., Mukhtar, M. S. & McDowell, J. M. Sparking a sulfur war between plants and pathogens. *Trends Plant. Sci.* **27**, 1253–1265. https://doi.org/10.1016/j.tplants.2022.07.007 (2022).
28. Thingujam, D., Pajerowska-Mukhtar, K. M. & Mukhtar, M. S. Duckweed: beyond an efficient plant model system. *Biomolecules* **14** https://doi.org/10.3390/biom14060628 (2024).

Acknowledgements

This work was supported by the National Science Foundation (IOS-2038872) to M.S.M. The authors gratefully acknowledge the resources provided by the University of Alabama at Birmingham IT-Research Computing group for high-performance computing (HPC) support and CPU time on the Cheaha compute cluster.

Author contributions

N.K. assembled the pipeline and wrote the manuscript. N.K. and M.S.M. initially conceptualized and refined the

protocol. N.K. conducted the initial analyses. M.S.M. and N.K. conceptualized the project. M.S.M. oversaw the analysis. All authors reviewed and edited the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91809-z.

Correspondence and requests for materials should be addressed to M.S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.